# A HIGH-COMPRESSION TRANSFORMER VAE FOR EFFICIENT 3D MEDICAL IMAGE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion Transformers have enabled large-scale image and video generation, but their extension to three-dimensional medical images is constrained by long spatial token sequences and the resulting attention cost. High spatial compression offers a practical route to volumetric generation, yet preserving fine anatomical and pathological detail at such compression remains challenging. We introduce a high-capacity Transformer variational autoencoder trained on BrainTumor-256K, a large multi-sequence brain tumor MRI collection. Our model compresses each spatial dimension by a factor of eight, yielding a Gaussian latent representation with $512\times$ fewer spatial locations. The model combines high reconstruction fidelity with efficient computation on this compact grid. Across four public brain MRI datasets, our model achieves the highest mean PSNR among the evaluated autoencoders, with consistent advantages in structural and perceptual metrics. Despite its large parameter count, it delivers the lowest encoding–decoding latency at $256^3$ among the evaluated autoencoders and requires substantially less training memory than comparably sized convolutional backbones. Trained exclusively on brain MRI, the model also reconstructs non-brain MRI and CT without target-domain fine-tuning, exceeding all evaluated medical autoencoders in PSNR on every non-brain dataset, and accommodates full volumes of varying sizes and shapes. Modality-controlled and tumor-mask-guided diffusion generation further demonstrate the utility of its compressed representation. These results establish high-capacity Transformer autoencoding as an effective approach to high-compression medical image representation and a practical foundation for volumetric generation. Code is available at `https://anonymous.4open.science/r/transformer-vae3d-2D31`.

## 1 INTRODUCTION

High spatial compression offers a practical route to scaling generative models to three-dimensional medical images. Latent diffusion reduces the cost of image synthesis by operating on compressed representations, while Diffusion Transformers make the spatial token count a key determinant of computational cost (Rombach et al., 2022; Peebles & Xie, 2023). For medical volumes, this challenge is amplified by the third spatial dimension: a $256^3$ image contains more than sixteen million voxels. Compressing such volumes must retain the anatomy and disease-related detail needed for faithful reconstruction. This creates a practical demand for volumetric autoencoders that combine strong spatial compression, high reconstruction fidelity and affordable computation during both training and inference.

Medical autoencoders already support brain MRI synthesis and controllable CT generation (Pinaya et al., 2022; Khader et al., 2023; Guo et al., 2025; Wang et al., 2025), but achieving high reconstruction fidelity and practical inference speed under strong spatial compression remains challenging. A compact latent grid must encode fine structures and lesions that occupy only a small fraction of the volume. Work on deep compression and transformer tokenizers motivates allocating greater model capacity to this representation problem (Chen et al., 2025; Hansen-Estruch et al., 2025). Transformer backbones offer a route to increasing capacity through their width and depth; in medical imaging, pursuing this route also calls for training data spanning anatomical, pathological and image-contrast variation. For volumetric CNNs, increasing feature width also increases the cost of activations pro-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

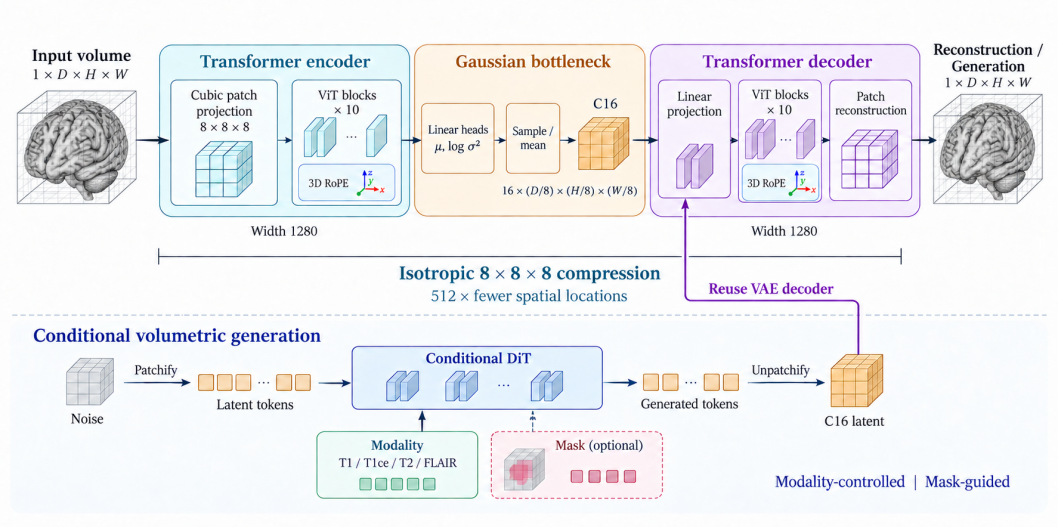


Figure 1: Volumetric Transformer VAE with isotropic $8 \times 8 \times 8$ compression and 3D RoPE. A conditional DiT generates latent tokens from noise using modality and optional tumor-mask conditions, then reuses the VAE decoder. Brain renderings are schematic.

cessed at high spatial resolutions. Moving repeated computation onto a compact token grid offers a different allocation of capacity. These considerations motivate combining a high-capacity Transformer VAE with large-scale brain MRI training.

We develop a high-capacity volumetric Gaussian VAE trained on BrainTumor-256K, a large multi-sequence brain tumor MRI collection. Cubic patch projection reduces each spatial axis by eight before the Transformer backbone, concentrating model capacity on a compact grid. A narrower latent variant further explores the trade-off between compression and image fidelity. Equal downsampling and bidirectional attention treat all three axes as spatial, while three-dimensional rotary position encoding supports inference beyond the training crop.

Our contributions are threefold:

1. **High-compression volumetric representation.** We introduce a high-capacity Transformer VAE trained on BrainTumor-256K, a large multi-sequence brain tumor MRI collection. The model combines isotropic eightfold compression along each spatial axis with a compact Gaussian latent representation, reducing spatial locations by a factor of 512 while preserving high image fidelity.
2. **High-fidelity medical autoencoding with efficient computation.** Our model achieves the highest mean PSNR across four public brain MRI datasets and the lowest encoding–decoding latency at $256^3$ among the evaluated autoencoders. Controlled resource measurements further show substantially lower training-memory requirements than convolutional VAEs of comparable parameter counts, demonstrating that large model capacity can remain computationally practical.
3. **A practical latent representation for volumetric generation.** We demonstrate full-volume brain MRI synthesis with a 2B-parameter Diffusion Transformer operating in the learned latent space, supporting both MRI-sequence control and tumor-mask guidance. Downstream resource measurements quantify the computational benefits of the compact spatial representation for DiT training.

## 2 Related Work

**High-compression image representations.** Variational and vector-quantized autoencoders provide continuous and discrete representations (Kingma & Welling, 2014; van den Oord et al., 2017). Perceptual and adversarial objectives improve reconstruction fidelity under compression (Esser

et al., 2021), while latent diffusion and Diffusion Transformers make representation size central to generative computation (Rombach et al., 2022; Peebles & Xie, 2023). DC-AE studies deep spatial compression, ViTok investigates Transformer tokenizer scaling, and TC-AE examines token capacity (Chen et al., 2025; Hansen-Estruch et al., 2025; Li et al., 2026b). Representation and prior-aligned autoencoders further study latent properties relevant to generation (Zheng et al., 2026; Yue et al., 2026). Our work brings high-capacity Transformer autoencoding to medical volumes, where reconstruction fidelity must be balanced against the cost of processing three spatial dimensions.

**Volumetric medical autoencoders.** Brain latent diffusion and Medical Diffusion established volumetric MRI and CT generation (Pinaya et al., 2022; Khader et al., 2023). MAISI combines volumetric compression with controllable CT synthesis, and 3D MedDiffusion uses a patch-volume autoencoder (Guo et al., 2025; Wang et al., 2025). MedVAE studies medical autoencoding, while NeuroQuant explores modality-aware anatomical VQ representations (Varma et al., 2026; Li et al., 2026a). CONFLUX and whole-volume translation further demonstrate volumetric latent modeling (Van Puyvelde et al., 2026; Molino et al., 2026). CONFLUX additionally provides a chest-CT CNN VAE with the same f8/C16 latent interface as our model, enabling a direct comparison of reconstruction and computational cost at matched compression.

**Video autoencoders for volumetric data.** CogVideoX, HunyuanVideo and Wan provide pretrained video VAEs that can be adapted to medical volumes (Yang et al., 2025; Kong et al., 2024; Team Wan et al., 2025). Their temporal–spatial designs contrast with our equal-axis compression and bidirectional processing. We include released video VAEs as reconstruction and latency baselines.

# 3 Volumetric Transformer Variational Autoencoding

## 3.1 Representation and compression

For a grayscale volume $x \in \mathbb{R}^{1\times D\times H\times W}$ with dimensions divisible by eight, the encoder produces a diagonal Gaussian posterior

$$q_\phi(z \mid x) = \mathcal{N}\big(\mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))\big), \qquad z \in \mathbb{R}^{16\times D/8\times H/8\times W/8}. \tag{1}$$

Training samples $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$; reconstruction evaluation uses the posterior mean. An ordinary 3D convolution with kernel and stride $(8, 8, 8)$ projects voxel patches into spatial tokens, following the patch-embedding approach of ViT (Dosovitskiy et al., 2021). Global transformer blocks process this compact grid, and the decoder projects the decoded tokens back to voxel patches (Figure 1).

We distinguish the spatial reduction from the scalar-element reduction:

$$R_{\text{grid}} = 8^3 = 512, \qquad R_{\text{elements}} = \frac{1 \cdot DHW}{16 \cdot (D/8)(H/8)(W/8)} = 32. \tag{2}$$

The primary model uses C16. We also evaluate a C8 variant with the same Transformer backbone and spatial downsampling; it increases $R_{\text{elements}}$ from 32 to 64 while leaving $R_{\text{grid}} = 512$ unchanged. At $256^3$, either latent contains $32^3$ spatial positions. Appendix B compares latent tensor sizes across interfaces; these ratios describe dimensional reduction rather than encoded bitrate.

## 3.2 Isotropic spatial compression

Equal downsampling preserves the aspect ratio of the spatial grid: input spacing $(s_D, s_H, s_W)$ becomes latent spacing $(8s_D, 8s_H, 8s_W)$. Isotropic inputs therefore retain an isotropic latent grid. Together with bidirectional attention, this design treats all three axes as spatial throughout encoding and decoding.

At equal C16 bottlenecks, $(8, 8, 8)$ compression produces half as many latent positions as $(4, 8, 8)$ compression for divisible input dimensions. At a fixed downstream DiT patch size, this halves the token count and quarters the number of dense-attention interactions. Appendix B gives the accounting, and Section 5.6 examines the resulting training cost.

### 3.3 TRANSFORMER BACKBONE AND VARIABLE-SHAPE INFERENCE

The backbone uses ten encoder and ten decoder blocks of width 1280, with sixteen attention heads, RMS normalization and SwiGLU feed-forward layers. Three-dimensional rotary position encoding applies query–key rotations along the spatial axes (Su et al., 2024; Heo et al., 2024), using coordinates derived from the current token grid. This allows the positional representation to follow the input geometry. Appendix D provides the architectural settings.

Training uses $128^3$ crops. At inference, volumes are padded to multiples of eight and cropped back to their original shape after decoding. The model therefore accepts dimensions and aspect ratios beyond the training crop; Appendix F summarizes the evaluated input shapes.

### 3.4 TRAINING OBJECTIVE

The training objective combines voxel reconstruction, perceptual supervision, a Gaussian prior penalty, discriminator feature matching, and adversarial training:

$$\mathcal{L}_{\text{rec}} = \frac{\lambda_1}{N}\|x - \hat{x}\|_1 + \lambda_p \mathcal{L}_{\text{perc}}(x, \hat{x}), \tag{3}$$

$$\mathcal{L}_G = \mathcal{L}_{\text{rec}} + \beta(t)\overline{D}_{\text{KL}}(q_\phi(z \mid x)\|\mathcal{N}(0, I)) + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{adv}}(t)a(t)\mathcal{L}_{\text{adv}}, \tag{4}$$

$$\mathcal{L}_{\text{fm}} = \sum_\ell \text{mean}\left|D_\ell(x) - D_\ell(\hat{x})\right|. \tag{5}$$

Here $N = DHW$, and the voxel-mean L1 term is averaged over the batch. $\overline{D}_{\text{KL}}$ is the mean scalar Gaussian KL divergence over spatial tokens and channels. Perceptual supervision uses a pretrained SqueezeNet on sampled slices from all three spatial axes (Zhang et al., 2018). Feature matching compares discriminator activations $D_\ell$ for original and reconstructed inputs, while $a(t)$ balances reconstruction and adversarial gradient magnitudes. Appendix A specifies the loss weights, sampling and regularization settings.

The KL and adversarial terms are enabled on warm-up schedules. We train with AdamW and cosine learning-rate decay on eight NVIDIA H200 GPUs with BF16 mixed precision, an effective batch size of 128 crops, and spatial and intensity augmentation (Appendix A).

## 4 EXPERIMENTAL SETUP

**Training collection and split.** BrainTumor-256K is a private multi-sequence brain tumor MRI collection containing approximately 256,000 volumes from 8,000 patients. It contains no volumes from the public evaluation datasets and is partitioned by patient into training and validation sets; Appendix A provides cohort statistics and training settings.

We evaluate two brain-trained VAE variants: C16 is the primary model, and C8 tests a narrower latent bottleneck. Both share the same architecture and training hyperparameters apart from the latent channel count. Efficiency and tumor-preservation measurements use C16 unless specified otherwise.

**Evaluation data and preprocessing.** We assess brain MRI autoencoding fidelity by comparing input volumes with their decoded reconstructions on BraTS, pooled OASIS-1/OASIS-2, IXI and MSD Brain (Marcus et al., 2007; 2010; Brain Development, n.d.; Antonelli et al., 2022). BraTS refers to the ISBI 2024 BraTS-GoAT generalizability subset (Conte et al., 2024). We also report results on the internal validation cohort, whose patients are disjoint from the training set. Cross-domain evaluations cover AMOS MRI, fastMRI knee, MSD heart and prostate MRI, and liver, lung, pancreas and kidney CT (Knoll et al., 2020; Ji et al., 2022; Antonelli et al., 2022; Heller et al., 2021). Volumes are reoriented, resampled to 1 mm spacing and percentile-normalized, and reconstructions are scored on a common reference grid. Appendix D specifies preprocessing and model-specific adaptations.

**Metrics and coverage.** We use PSNR as the primary reconstruction metric and report MS-SSIM, SSIM and LPIPS alongside it (Wang et al., 2004; 2003; Zhang et al., 2018), with structural metrics

Table 1: Brain MRI autoencoding fidelity (mean PSNR in dB ↑) and latent compression. Rows are ordered by PSNR on the internal BrainTumor-256K validation cohort.

| Model | BraTS | OASIS | IXI | MSD Brain | BrainTumor-256K | Compression | Spatial Reduction | Channel | Element Reduction |
|---|---|---|---|---|---|---|---|---|---|
| Ours (C16) | **31.79** | **29.46** | **31.81** | **30.91** | **29.09** | $8 \times 8 \times 8$ | 512× | 16 | 32× |
| CogVideoX | 20.95 | 28.71 | 24.46 | 20.60 | 28.91 | $4 \times 8 \times 8$ | 256× | 16 | 16× |
| HunyuanVideo | 30.66 | 28.72 | 30.79 | 29.84 | 28.78 | $4 \times 8 \times 8$ | 256× | 16 | 16× |
| Wan | 29.98 | 28.20 | 30.30 | 29.36 | 28.04 | $4 \times 8 \times 8$ | 256× | 16 | 16× |
| MAISI-MR | 27.68 | 26.12 | 27.75 | 27.61 | 26.81 | $4 \times 4 \times 4$ | 64× | 4 | 16× |
| Ours (C8) | 28.59 | 26.95 | 28.47 | 27.88 | 26.56 | $8 \times 8 \times 8$ | 512× | 8 | 64× |
| MedVAE-f4c1 | 28.07 | 25.67 | 27.85 | 27.56 | 26.16 | $4 \times 4 \times 4$ | 64× | 1 | 64× |
| MAISI-CT | 26.12 | 25.39 | 25.54 | 26.21 | 26.16 | $4 \times 4 \times 4$ | 64× | 4 | 16× |
| Patch-Volume | 22.85 | 22.01 | 22.80 | 22.74 | 22.03 | $8 \times 8 \times 8$ | 512× | 8 | 64× |
| MedVAE-f8c1 | 22.88 | 20.23 | 21.88 | 22.94 | 21.91 | $8 \times 8 \times 8$ | 512× | 1 | 512× |
| Brain-LDM | 18.61 | 17.88 | 18.20 | 18.55 | 21.10 | $8 \times 8 \times 8$ | 512× | 3 | 170.7× |
| CONFLUX | 21.86 | 14.98 | 24.10 | 21.60 | 19.93 | $8 \times 8 \times 8$ | 512× | 16 | 32× |

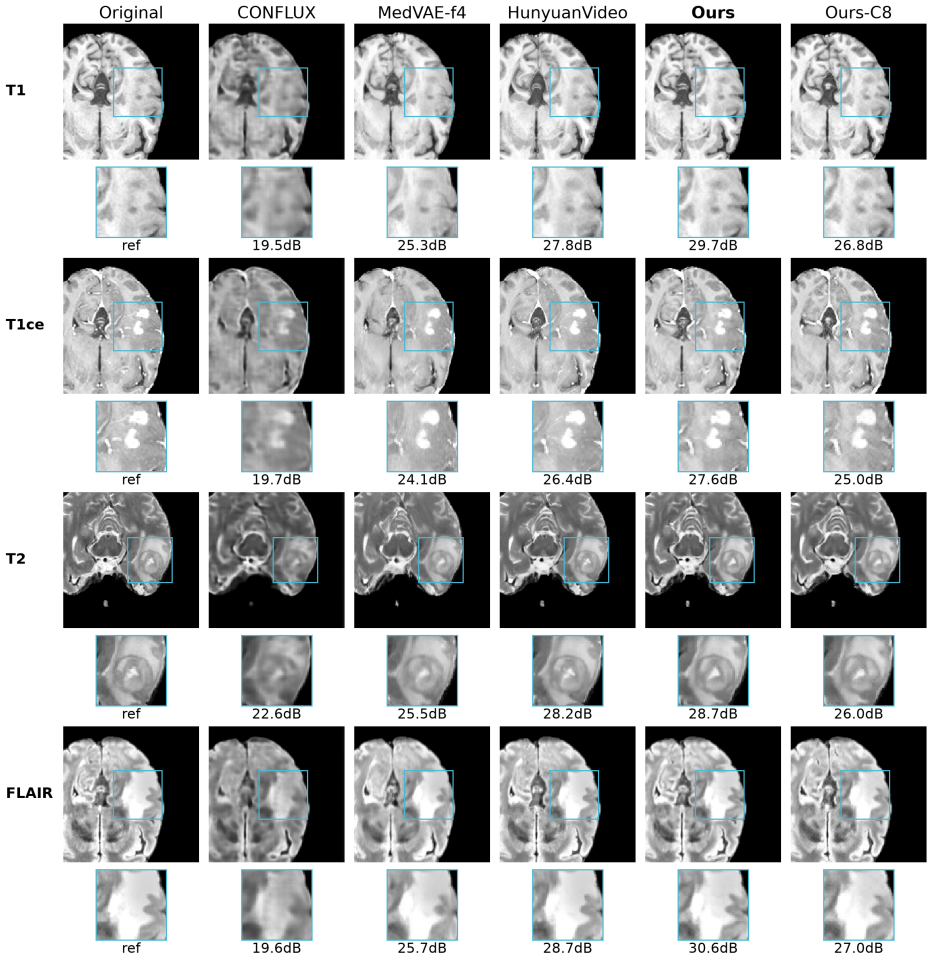


Figure 2: Brain MRI inputs and decoded reconstructions across four sequences, with shared views and display limits. Each row pair shows an axial slice and a zoomed region centered on the tumor annotation; the number below each column is the PSNR (dB) of that reconstruction over the whole volume, not only the zoomed region. Ours and Ours-C8 denote C16 and C8; MedVAE-f4 denotes MedVAE-f4c1.

evaluated within a foreground bounding box. All models are compared on the same valid volumes within each dataset. Paired confidence intervals use subject-grouped bootstrap resampling to account for multiple sequences and examinations per patient. Metric definitions, statistical procedures and evaluation coverage are provided in Appendices D, E and G.

**Baselines and timing.** We compare released medical autoencoders and three video VAEs at their native compression rates (Table 1; Appendix C). Encoding–decoding latency is measured on one RTX PRO 6000 Blackwell Server Edition GPU using fixed-size synthetic volumes and batch size one. All models are benchmarked using 16-bit inference (BF16 or FP16, depending on the implementation). Appendix D specifies timing boundaries and model-specific inference policies.

# 5 Results

## 5.1 Brain MRI autoencoding

Our C16 model achieves the highest mean PSNR among the evaluated autoencoders on all four public brain MRI evaluation sets (Table 1), and also leads in MS-SSIM, SSIM and LPIPS on these datasets. The comparison includes lower-compression alternatives and the f8/C16 CONFLUX VAE, a chest-CT model not expected to excel on brain MRI (see Section 5.4 for CT). On the internal cohort, C16 also leads in PSNR and MS-SSIM, while HunyuanVideo and Wan lead in SSIM and LPIPS, respectively (Table 11). Figure 2 and Appendix L show example reconstructions; Appendices G and E report all metrics and paired PSNR differences.

**Increasing element compression.** The C8 variant doubles element compression while retaining the same spatial grid. Although its PSNR is lower than C16, it remains higher than MedVAE-f4c1 and Patch-Volume at the same $64\times$ element reduction, and higher than MAISI-MR at $16\times$, across the public brain datasets. The high-capacity backbone thus supports a more compact latent representation, while C16 offers the strongest reconstruction fidelity.

## 5.2 Encoding and decoding efficiency

At $256^3$, our model achieves the lowest encoding–decoding latency among the evaluated autoencoders despite having the largest parameter count (Table 2). It is $4.95\times$ faster than CONFLUX, which uses the same f8/C16 latent interface. Together with the brain MRI autoencoding results, this demonstrates a favorable balance of fidelity and inference cost (Figure 4).

Because global attention cost grows quadratically with token count, the advantage narrows at $384^3$: our model remains faster than the video VAEs, but Brain-LDM and both MAISI variants are faster, and CONFLUX runs out of memory.

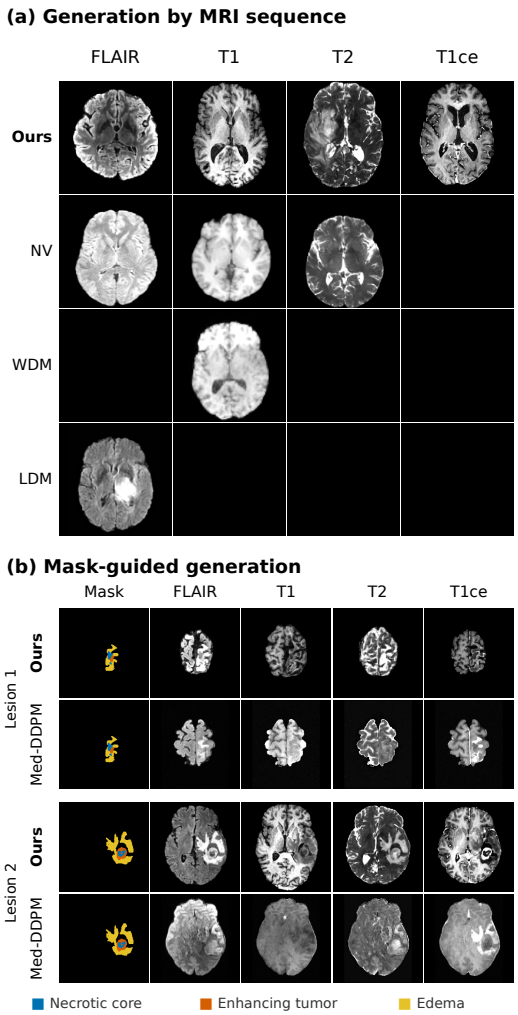


Figure 3: Generation examples: (a) sequence control (NV: NV-Generate-MR-Brain; LDM: MONAI BraTS-LDM); (b) Ours and Med-DDPM for two lesions with shared masks. Black: unsupported.

Table 2: Encoding–decoding latency (ms; lower is better). Spatial Reduction denotes the reduction in grid locations; the final column reports speedup relative to each baseline at $256^3$. OOM indicates insufficient GPU memory.

| Model | Params (M) | Spatial Reduction | $128^3$ (ms) | $256^3$ (ms) | $384^3$ (ms) | Speedup at $256^3$ |
|---|---|---|---|---|---|---|
| Ours | 525.8 | 512× | 36.7 | 653.0 | 5451.9 | 1.00× |
| Wan | 126.9 | 256× | 303.6 | 1941.3 | 7669.7 | 2.97× |
| CogVideoX | 215.6 | 256× | 471.2 | 4146.7 | 14653.4 | 6.35× |
| HunyuanVideo | 246.5 | 256× | 471.1 | 4378.8 | 15060.4 | 6.71× |
| MedVAE-f4c1 | 161.0 | 64× | 623.3 | 8407.3 | 31112.9 | 12.88× |
| MedVAE-f8c1 | 246.0 | 512× | 519.5 | 8129.0 | 28437.9 | 12.45× |
| MAISI-MR | 20.9 | 64× | 223.4 | 1445.6 | 4794.9 | 2.21× |
| MAISI-CT | 20.9 | 64× | 221.0 | 1431.4 | 4834.9 | 2.19× |
| Brain-LDM | 13.8 | 512× | 122.4 | 1017.0 | 3453.1 | 1.56× |
| Patch-Volume | 34.8 | 512× | 132.2 | 1081.1 | 51880.2 | 1.66× |
| CONFLUX | 246.5 | 512× | 384.2 | 3230.3 | OOM | 4.95× |

### 5.3 Training efficiency at comparable capacity

To compare training cost at similar capacities, we scale the width of Transformer and CONFLUX backbones while retaining the f8/C16 interface. Under a shared MSE objective, the Transformer requires approximately 82–87% less peak memory at batch size one across the tested capacities, allowing substantially larger training batches (Table 3).

Our model also completes a training step $6.0\times$ faster than the released CONFLUX VAE at batch size one, despite its larger parameter count. Appendix D describes the measurement protocol.

Table 3: Training memory at $128^3$ on one H200 under a shared MSE objective. B1 denotes batch size one; the final column gives the largest feasible batch among tested powers of two.

| Backbone | Params (M) | B1 peak (GiB) | Largest tested batch |
|---|---|---|---|
| ViT-VAE (width 880) | 249.0 | 6.4 | 16 |
| CNN-VAE (base 128) | 246.5 | 49.2 | 2 |
| ViT-VAE (width 1120) | 403.0 | 9.6 | 16 |
| CNN-VAE (base 160) | 384.9 | 57.1 | 2 |
| ViT-VAE (width 1280; ours) | 525.8 | 12.6 | 16 |
| CNN-VAE (base 192) | 554.0 | 69.0 | 1 |

### 5.4 Variable-shape and cross-domain autoencoding

Trained exclusively on brain MRI, our model is evaluated on non-brain MRI and CT without any target-domain fine-tuning. These evaluations span both new imaging domains and input sizes and aspect ratios beyond the $128^3$ training crop. Across all eight non-brain MRI and CT datasets, C16 ranks within the top three in PSNR, exceeding every medical autoencoder by 0.7–4.6 dB, including the CT-trained MAISI-CT and CONFLUX on all four CT datasets, and trailing only the HunyuanVideo and Wan video VAEs, which operate at lower spatial compression (Table 16). HunyuanVideo leads most non-brain comparisons, while our model achieves the lowest mean LPIPS on prostate MRI. Appendices F and G report shape coverage and the complete matched-volume comparisons.

### 5.5 Tumor-region preservation

We assess tumor-region preservation by applying the same frozen NS-HGlio segmenter (Abayazeed et al., 2023) to original inputs and reconstructions from our model and HunyuanVideo. The paired evaluation uses four-sequence examinations from preoperative BraTS-GLI validation patients, with ground-truth annotations used only for scoring. Appendix H details the protocol.

Our reconstructions retain tumor segmentation performance close to that of the original inputs, with mean Dice decreasing by less than one percentage point in each tumor region (Table 4). Compared with HunyuanVideo, our model has higher TC and ET Dice, while HunyuanVideo has higher WT Dice. Paired confidence intervals support the TC and ET gains (Table 20).

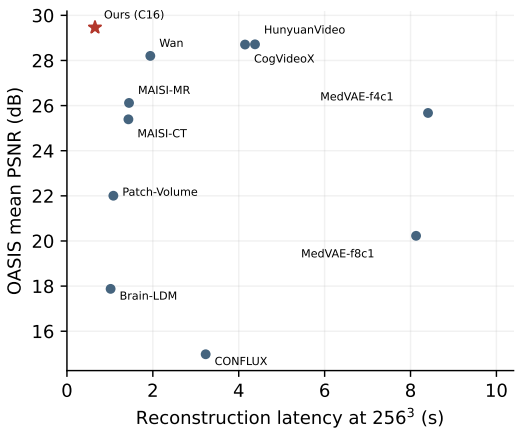


Figure 4: OASIS mean PSNR versus encoding–decoding latency at $256^3$; latency is measured on fixed-size synthetic inputs.

Table 4: Tumor segmentation Dice (%) on BraTS-GLI validation inputs; WT/TC/ET: whole tumor, tumor core, enhancing tumor.

| Input to segmenter | WT | TC | ET |
|---|---|---|---|
| Original | 83.26 | 84.92 | 82.98 |
| Ours (C16) | 83.02 | 84.46 | 82.25 |
| HunyuanVideo | 84.52 | 82.88 | 80.68 |

### 5.6 DiT training resources

We assess the downstream cost of compressed representations by training DiTs of three sizes on synthetic latent tensors. At a fixed latent patch size, the f8/C16 interface reduces the 2B-parameter DiT training-step time by $1.4\times$ relative to the $(4, 8, 8)$/C16 interface of video VAEs, and by $16\times$ relative to the f4/C4 interface of MAISI, while approximately halving peak memory in the latter comparison. Appendix I gives the full measurements and protocol.

Increasing the DiT patch size for f4/C4 from two to four equalizes the token count and yields resource use similar to f8/C16. Cost is thus set jointly by VAE compression and DiT patchification; the difference is whether spatial reduction is performed by a learned reconstruction encoder or by a single linear patch projection inside the generator.

### 5.7 Conditional volumetric generation

We demonstrate generation from noise using a 2B-parameter multimodal DiT in the C16 latent space. Modality-controlled synthesis conditions on the requested MRI sequence; mask-guided synthesis additionally conditions on a tumor mask. Both modes generate complete T1, contrast-enhanced T1, T2 and FLAIR volumes (Figure 3). We evaluate distributional fidelity using slice-FID (Heusel et al., 2017) (Table 5). Because FID depends on the reference distribution and each generator reflects its own training data, we report two reference cohorts: internal validation images, which share acquisition characteristics with our training data, and BraTS, an external public cohort that also provides the training data of MONAI BraTS-LDM and WDM. The released modality-controlled baselines each support only a subset of sequences, and none generates contrast-enhanced T1. Against internal references, our model achieves the lowest slice-FID for T1 and FLAIR, while NV-Generate is lower for T2; against BraTS references, the BraTS-trained MONAI BraTS-LDM is lower on FLAIR, NV-Generate is lower on its three supported sequences, likely reflecting its broader training distribution of approximately 28,000 multi-site brain MRI scans, and our model remains ahead of WDM on T1. In mask-guided generation, where both models receive the same tumor masks, our model achieves lower slice-FID than Med-DDPM (Dorjsembe et al., 2024) on all four sequences against both reference cohorts. A frozen NS-HGlio segmenter recovers the conditioning tumors in our generated volumes with a Dice of 0.82 and a recall of 0.91, although our model generates each sequence independently while Med-DDPM additionally conditions on a shared brain-region label (Table 22). Five additional paired examples illustrate the generated MRI appearance under shared mask conditions (Figures 6 and 7). Appendix J details the evaluation.

## 6 Discussion and Limitations

The results show that a high-capacity Transformer VAE can combine strong spatial compression, high-fidelity brain MRI autoencoding and efficient computation. Without target-domain fine-tuning,

Table 5: Slice-FID ↓ by generation task and reference cohort. Internal and BraTS denote original-image reference cohorts. Dashes indicate unsupported sequences.

| Model | Reference | T1 | T1ce | T2 | FLAIR |
|---|---|---|---|---|---|
| *Modality-controlled* | | | | | |
| NV-Generate | Internal | 47.73 | – | 44.19 | 48.20 |
| NV-Generate | BraTS | 66.80 | – | 34.12 | 38.88 |
| MONAI BraTS-LDM | Internal | – | – | – | 63.84 |
| MONAI BraTS-LDM | BraTS | – | – | – | 51.66 |
| WDM | Internal | 123.36 | – | – | – |
| WDM | BraTS | 102.34 | – | – | – |
| Ours | Internal | 46.74 | 34.99 | 48.17 | 29.70 |
| Ours | BraTS | 77.28 | 60.85 | 54.93 | 56.47 |
| *Mask-guided* | | | | | |
| Med-DDPM | Internal | 156.87 | 95.36 | 104.75 | 96.64 |
| Med-DDPM | BraTS | 111.83 | 91.83 | 105.19 | 80.52 |
| Ours | Internal | 45.75 | 32.58 | 42.57 | 29.15 |
| Ours | BraTS | 76.89 | 61.27 | 52.09 | 56.89 |

it also exceeds all evaluated medical autoencoders, including MAISI-CT, on non-brain MRI and CT, suggesting that the learned representation extends well beyond the training domain, although fidelity still varies by anatomy and modality. Fast encoding could also enable on-the-fly latent encoding during diffusion training.

Our reconstruction results reflect model capacity and training scale jointly, and we view the two as complementary: large, diverse training data supplies anatomical, pathological and contrast variation, while the Transformer backbone makes a 526M-parameter volumetric autoencoder affordable to train and fast at inference (Sections 5.2–5.3). Separating their contributions through matched training runs would require retraining several large volumetric autoencoders on the same corpus, which we leave to future work. The autoencoder resource benchmark measures per-step computation and memory under a shared MSE objective; it excludes perceptual and adversarial losses and does not measure total training cost. The quadratic cost of global attention also narrows the latency advantage for larger volumes.

Private training data limit exact replication, and we therefore treat the public datasets as the primary reconstruction benchmarks. Tumor preservation is assessed with one frozen segmenter, and normalized CT reconstruction does not establish calibrated Hounsfield-unit accuracy or clinical utility. Slice-FID assesses distributional similarity, while the segmentation-based evaluation measures tumor-mask adherence; neither fully characterizes three-dimensional anatomical consistency.

## 7 Conclusion

We present a high-capacity Transformer VAE that combines isotropic spatial compression with high-fidelity brain MRI autoencoding and efficient computation. A narrower latent bottleneck offers greater element compression, while the learned representation supports modality-controlled and tumor-mask-guided volumetric synthesis. These results make high-capacity Transformer VAEs a practical foundation for medical image representation and generation, and cross-domain results that surpass the evaluated medical autoencoders from brain MRI training alone suggest that extending training to additional anatomies and imaging modalities is a promising path toward a general-purpose medical VAE.

## Reproducibility Statement

We describe the architecture, training settings, compression accounting, evaluation preprocessing, metrics, paired statistics and timing boundaries. Code is provided at `https://anonymous.4open.science/r/transformer-vae3d-2D31`. VAE weights will be publicly released upon publication. The DiT weights will not be released because unconditional generative models

can memorize and reproduce training images, whereas the VAE only reconstructs volumes supplied by the user. The private training corpus and subject-level records will not be distributed.

## ETHICS STATEMENT

The study was approved by an institutional review board. Institutional identifiers and the approval number are omitted for double-blind review. The internal MRI data and subject-level records are private and will not be released. The model is intended for research; clinical use would require further validation.

## AI USE STATEMENT

Generative AI tools assisted with literature search, experiment planning, benchmark code development, analysis of results, table preparation, and manuscript writing and editing. The authors verified all code, results and text and take full responsibility for the content. Figure 1 was generated using an AI image-generation tool from an author-specified description of the model architecture.

## REFERENCES

Aly H. Abayazeed, Ahmed Abbassy, Michael Müeller, et al. NS-HGlio: A generalizable and repeatable HGG segmentation and volumetric measurement AI algorithm for the longitudinal MRI assessment to inform RANO in trials and clinics. *Neuro-Oncology Advances*, 5(1):vdac184, 2023. doi: 10.1093/noajnl/vdac184. URL https://doi.org/10.1093/noajnl/vdac184.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, et al. The Medical Segmentation Decathlon. *Nature Communications*, 13:4128, 2022. doi: 10.1038/s41467-022-30695-9. URL https://www.nature.com/articles/s41467-022-30695-9.

Brain Development. IXI dataset. https://brain-development.org/ixi-dataset/, n.d. Dataset resource; accessed September 22, 2026.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wH8XXUOUZU.

Gian Marco Conte, Ujjwal Baid, Spyridon Bakas, Mariam Aboian, Maruf Adewole, Jake Albrecht, Udunna Anazodo, Evan Calabrese, Verena Chung, Anastasia Janas, et al. BraTS Generalizability Across Tumors, 2024. URL https://doi.org/10.5281/zenodo.10687238. ISBI 2024 challenge description.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024. doi: 10.1109/JBHI.2024.3385504. URL https://ieeexplore.ieee.org/document/10493074.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

Paul Friedrich, Julia Wolleb, Florentin Bieder, Alicia Durrer, and Philippe C. Cattin. WDM: 3D Wavelet Diffusion Models for High-Resolution Medical Image Synthesis. In *Deep Generative Models: DGM4MICCAI 2024*, pp. 11–21. Springer, 2024. doi: 10.1007/978-3-031-72744-3_2. URL https://link.springer.com/chapter/10.1007/978-3-031-72744-3_2.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI: Medical AI for Synthetic Imaging. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 4430–4441, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Guo_MAISI_Medical_AI_for_Synthetic_Imaging_WACV_2025_paper.html.

Philippe Hansen-Estruch, David Yan, Ching-Yao Chuang, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from Scaling Visual Tokenizers for Reconstruction and Generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22023–22043. PMLR, 2025. URL https://proceedings.mlr.press/v267/hansen-estruch25a.html.

Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical Image Analysis*, 67:101821, 2021. doi: 10.1016/j.media.2020.101821. URL https://www.sciencedirect.com/science/article/abs/pii/S1361841520301857.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary Position Embedding for Vision Transformer. In *Computer Vision – ECCV 2024*. Springer, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1584_ECCV_2024_paper.php.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5MB model size. arXiv preprint arXiv:1602.07360, 2016. URL https://arxiv.org/abs/1602.07360.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ee604e1bedbd069d9fc9328b7b9584be-Paper-Datasets_and_Benchmarks.pdf.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13:7303, 2023. doi: 10.1038/s41598-023-34341-2. URL https://www.nature.com/articles/s41598-023-34341-2.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, et al. fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020. doi: 10.1148/ryai.2020190007. URL https://pubs.rsna.org/doi/abs/10.1148/ryai.2020190007.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv preprint arXiv:2412.03603, 2024. URL https://arxiv.org/abs/2412.03603.

Mingjie Li, Edward Kim, Yue Zhao, Ehsan Adeli, and Kilian M. Pohl. Modality-Aware and Anatomical Vector-Quantized Autoencoding for Multimodal Brain MRI. arXiv preprint arXiv:2604.05171, 2026a. URL https://arxiv.org/abs/2604.05171.

Teng Li, Ziyuan Huang, Cong Chen, Yangfu Li, Yuanhuiyi Lyu, Dandan Zheng, Chunhua Shen, and Jun Zhang. TC-AE: Unlocking Token Capacity for Deep Compression Autoencoders. arXiv preprint arXiv:2604.07340, 2026b. URL https://arxiv.org/abs/2604.07340.

Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007. doi: 10.1162/jocn.2007.19.9.1498.

Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies: Longitudinal mri data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010. doi: 10.1162/jocn.2009.21407.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694. URL https://doi.org/10.1109/TMI.2014.2377694.

Daniele Molino, Alessio Zoboli, Camillo Maria Caruso, Valerio Guarrasi, and Paolo Soda. Compositional Cross-Modality Translation via Whole-Volume Multitask Latent Flow Matching. arXiv preprint arXiv:2608.08135, 2026. URL https://arxiv.org/abs/2608.08135.

MONAI Consortium. Brain Image Synthesis Latent Diffusion Model. MONAI Model Zoo bundle, n.d.a. URL https://github.com/Project-MONAI/model-zoo/tree/dev/models/brain_image_synthesis_latent_diffusion_model. Model documentation; accessed September 25, 2026.

MONAI Consortium. BraTS MRI Generative Diffusion. MONAI Model Zoo bundle, n.d.b. URL https://github.com/Project-MONAI/model-zoo/tree/dev/models/brats_mri_generative_diffusion. Model documentation; accessed September 25, 2026.

NVIDIA. NV-Generate-MR. Hugging Face model release, 2025. URL https://huggingface.co/nvidia/NV-Generate-MR. Model card; accessed September 25, 2026.

NVIDIA. NV-Generate-MR-Brain. Hugging Face model release, 2026. URL https://huggingface.co/nvidia/NV-Generate-MR-Brain. Model card; accessed September 25, 2026.

NVIDIA. NV-Generate-CT. Hugging Face model release, n.d. URL https://huggingface.co/nvidia/NV-Generate-CT. Model card; accessed September 25, 2026.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.pdf.

Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain Imaging Generation with Latent Diffusion Models. In *Deep Generative Models: DGM4MICCAI 2022*, pp. 117–126. Springer, 2022. doi: 10.1007/978-3-031-18576-2_12. URL https://link.springer.com/chapter/10.1007/978-3-031-18576-2_12.

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*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Noam Shazeer. GLU Variants Improve Transformer. arXiv preprint arXiv:2002.05202, 2020. URL https://arxiv.org/abs/2002.05202.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, et al. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint arXiv:2503.20314, 2025. URL https://arxiv.org/abs/2503.20314.

Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010. doi: 10.1109/TMI.2010.2046908. URL https://doi.org/10.1109/TMI.2010.2046908.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Max Van Puyvelde, Halil Ibrahim Gulluk, Wim Van Criekinge, and Olivier Gevaert. CONFLUX: A Latent Diffusion Model for 3D Chest-CT Synthesis with RL Post-Training. arXiv preprint arXiv:2607.02998, 2026. URL https://arxiv.org/abs/2607.02998.

Maya Varma, Ashwin Kumar, Rogier van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Joseph Marcel Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay S. Chaudhari. MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders. In *Proceedings of The 8th International Conference on Medical Imaging with Deep Learning*, volume 301 of *Proceedings of Machine Learning Research*, pp. 1597–1626. PMLR, 2026. URL https://proceedings.mlr.press/v301/varma26a.html.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3D MedDiffusion: A 3D Medical Latent Diffusion Model for Controllable and High-quality Medical Image Generation. *IEEE Transactions on Medical Imaging*, 44(12):4960–4972, 2025. doi: 10.1109/TMI.2025.3585372. URL https://doi.org/10.1109/TMI.2025.3585372.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems and Computers*, 2003. doi: 10.1109/ACSSC.2003.1292216. URL https://ieeexplore.ieee.org/abstract/document/1292216.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ece.uwaterloo.ca/~z70wang/research/ssim/.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, and Yali Wang. What Matters for Diffusion-Friendly Latent Manifold? Prior-Aligned Autoencoders for Latent Diffusion. arXiv preprint arXiv:2605.07915, 2026. URL https://arxiv.org/abs/2605.07915.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Can Zhao, Pengfei Guo, Dong Yang, Yufan He, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI-v2: Accelerated 3D High-Resolution Medical Image Synthesis with Rectified Flow and Region-specific Contrastive Loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13088–13098, 2026. doi: 10.1609/aaai.v40i15.38309. URL https://ojs.aaai.org/index.php/AAAI/article/view/38309.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

# A Training Settings

BrainTumor-256K contains 256,277 three-dimensional sequence entries from 8,098 patients across 71,942 examinations. The training and validation sets are split by patient. Table 6 summarizes the training configuration, which is shared by C16 and C8; checkpoints are selected by reconstruction performance on the validation set. The perceptual loss applies LPIPS with ImageNet-pretrained SqueezeNet-1.1 features (Iandola et al., 2016; Zhang et al., 2018) to slices sampled with shared indices from the original and reconstructed volumes, summed over the three axes. KL regularization is annealed over 20,000 iterations; the adversarial term is enabled after 2,000 iterations and ramps to its target over the next 2,000. The discriminator uses hinge loss with R1 regularization every 16 iterations. Augmentation applies random affine transforms and flips (probability 0.5 each) and gamma $\in [0.9, 1.1]$ and intensity shifts of magnitude 0.05 (probability 0.3 each) to $[0, 1]$-normalized intensities, which are then mapped to $[-1, 1]$.

# B Compression and DiT Token Accounting

For a downstream latent patch of $p_D \times p_H \times p_W$, the token count is

$$N_{\mathrm{DiT}} = \frac{DHW}{f_D f_H f_W p_D p_H p_W}, \tag{6}$$

assuming divisible dimensions and no additional boundary tokens. Halving $N$ halves token-linear costs and quarters the $N^2$ interaction count of dense attention. The channel count $C$ controls per-location feature width and the DiT input/output projections; spatial token count is determined by the grid and patch size.

# C Evaluated Autoencoder Interfaces

Video VAEs receive grayscale volumes replicated across their RGB input channels. Compression ratios are computed relative to the original single-channel volume. For Patch-Volume, each spatial position holds an eight-dimensional embedding associated with one codebook index; the table reports embedding-element compression.

Table 6: Training configuration.

| Setting | Value |
|---|---|
| Training hardware | 8 NVIDIA H200 GPUs |
| Training crop | $128 \times 128 \times 128$ |
| Precision | BF16 mixed precision |
| Batch per GPU / accumulation iterations | 8 / 2 |
| Effective generator batch | 128 crops |
| Optimizer | AdamW, $\beta = (0.9, 0.95)$ |
| Weight decay / gradient clipping | 0.01 / 1.0 |
| Initial / minimum learning rate | $5 \times 10^{-4}$ / $5 \times 10^{-5}$ |
| Learning-rate schedule | Cosine decay |
| L1 / perceptual / feature-matching weights | 1 / 1 / 1 |
| Perceptual backbone | SqueezeNet-1.1 (LPIPS) |
| Perceptual slice fraction per axis | 0.25 |
| Effective KL coefficient at evaluated checkpoint | $1 \times 10^{-6}$ |
| Target adversarial multiplier | 1, with adaptive gradient balancing |
| VAE EMA decay | 0.9995 |

Table 7: Analytical latent sizes for grayscale $256^3$ input, before extra DiT patchification and model-specific boundary padding. "C" denotes continuous latent channels.

| Reduction / channels | Spatial positions | Scalar elements | Grid ratio | Element ratio |
|---|---|---|---|---|
| $(8, 8, 8)$ / C16 | 32,768 | 524,288 | 512 | 32 |
| $(4, 8, 8)$ / C16 | 65,536 | 1,048,576 | 256 | 16 |
| $(8, 8, 8)$ / C8 | 32,768 | 262,144 | 512 | 64 |
| $(8, 8, 8)$ / C32 | 32,768 | 1,048,576 | 512 | 16 |
| $(8, 8, 8)$ / C1 | 32,768 | 32,768 | 512 | 512 |

## D Architecture and Evaluation Details

**Architecture.** The C16 VAE has 525,766,400 parameters; C8 has 525,735,680. The bottleneck projections change with channel count, while the Transformer blocks remain the same. Each of its ten encoder and ten decoder blocks uses width 1280 and sixteen attention heads of dimension 80, with SwiGLU expansion factor four (Zhang & Sennrich, 2019; Shazeer, 2020). Of the 80 channels per head, 78 receive rotary encoding, split equally across the three spatial axes, and two remain unrotated. Rotary frequencies use base 10,000; coordinates are computed from the input token grid.

**Evaluation preprocessing and metrics.** OASIS combines T1-weighted inputs from OASIS-1 and OASIS-2; IXI uses a multi-sequence subset. BraTS sequence labels, including native and contrast-enhanced T1, follow the dataset metadata. Reference volumes use RAS orientation, 1 mm resampling and normalization between the 0.5th and 99.5th intensity percentiles. CONFLUX uses the shared normalized reference for MRI and HU-based preprocessing with tiled inference for CT.

MS-SSIM uses the three-scale MONAI implementation with kernel seven and weights $(0.0448, 0.2856, 0.3001)$. Structural metrics use a foreground bounding box with four-voxel padding. Foreground extraction thresholds normalized intensity at 0.01 and applies morphological closing and hole filling. LPIPS uses a slice-based SqueezeNet implementation.

**Baseline implementations.** Brain-LDM uses the UK Biobank autoencoder released with MONAI's brain image synthesis bundle (Pinaya et al., 2022; MONAI Consortium, n.d.a). MAISI-MR and MAISI-CT denote the released NV-Generate-MR and NV-Generate-CT autoencoders, respectively (Guo et al., 2025; Zhao et al., 2026; NVIDIA, 2025; n.d.). Patch-Volume denotes the autoencoder from 3D MedDiffusion (Wang et al., 2025). MedVAE-f4c1 and MedVAE-f8c1 are the two evaluated MedVAE variants (Varma et al., 2026). We also evaluate CONFLUX (Van Puyvelde et al., 2026) and the released CogVideoX, HunyuanVideo and Wan VAEs (Yang et al., 2025; Kong et al., 2024; Team Wan et al., 2025).

Table 8: Native latent interfaces and parameter counts of the evaluated autoencoders. Ratios use a single-channel image and divisible dimensions. VQ channels describe embedding width, not independent codebook indices.

| Model | Params M | Latent type | Factors | $C$ | Grid ratio | Element ratio |
|---|---|---|---|---|---|---|
| Ours (C16) | 525.8 | Gaussian | 8/8/8 | 16 | 512 | 32.0 |
| Ours (C8) | 525.7 | Gaussian | 8/8/8 | 8 | 512 | 64.0 |
| HunyuanVideo | 246.5 | Gaussian | 4/8/8 | 16 | 256 | 16.0 |
| Wan | 126.9 | Gaussian | 4/8/8 | 16 | 256 | 16.0 |
| CogVideoX | 215.6 | Gaussian | 4/8/8 | 16 | 256 | 16.0 |
| MedVAE-f4c1 | 161.0 | Gaussian | 4/4/4 | 1 | 64 | 64.0 |
| MedVAE-f8c1 | 246.0 | Gaussian | 8/8/8 | 1 | 512 | 512.0 |
| Patch-Volume | 34.8 | VQ embedding | 8/8/8 | 8 | 512 | 64.0 |
| MAISI-MR | 20.9 | Gaussian | 4/4/4 | 4 | 64 | 16.0 |
| MAISI-CT | 20.9 | Gaussian | 4/4/4 | 4 | 64 | 16.0 |
| Brain-LDM | 13.8 | Gaussian | 8/8/8 | 3 | 512 | 170.7 |
| CONFLUX | 246.5 | Gaussian | 8/8/8 | 16 | 512 | 32.0 |

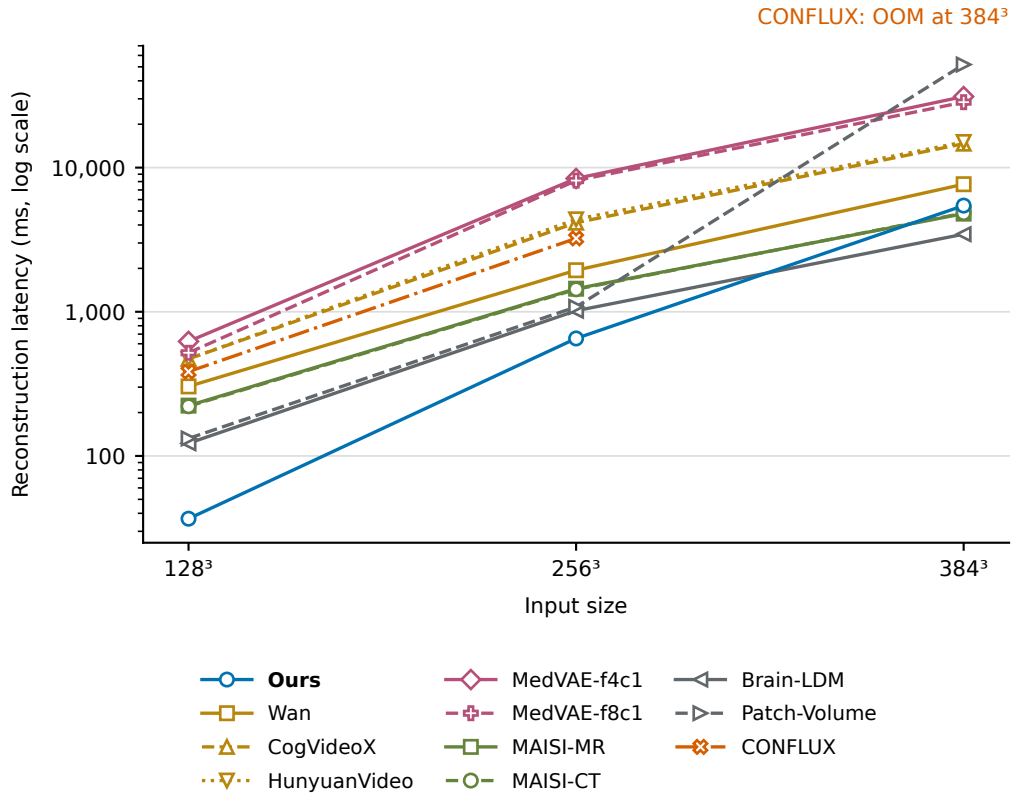


Figure 5: Encoding–decoding latency across input sizes from Table 2, with a logarithmic vertical axis. Points denote measured times; lines connect the three tested sizes. CONFLUX runs out of memory at $384^3$.

**Reconstruction timing.** Each input size uses two warmups and 25 measured reconstruction calls on fixed-size synthetic volumes, with batch size one. CUDA events enclose the full encoding–decoding call, including transfers, padding and model-specific tiling; disk loading and spatial resampling are excluded. The measurements follow the evaluated implementations' inference policies, including MedVAE's sliding-window path. CONFLUX timing uses whole-volume inference. Parameter counts include the autoencoder only. Speedup at $256^3$ is baseline latency divided by our latency. Figure 5 plots the timings across the three input sizes.

**Autoencoder training resources.** Memory comparisons use $128^3$ inputs on one H200, BF16 autocast with FP32 parameters, a shared MSE reconstruction objective and AdamW, without activation checkpointing. Width-scaled variants are randomly initialized. The largest feasible batch is selected among tested powers of two. A separate batch-one timing comparison measures complete forward, backward and AdamW steps, with three warmups and seven measured steps. Median times are 91 ms for our model and 546 ms for the released 246.5M CONFLUX VAE.

## E Paired Reconstruction Statistics

Table 9 uses the same volumes as the four public-data columns of the main brain comparison. Each of 2,000 bootstrap replicates resamples subjects with replacement within each dataset, retaining all of a subject's modalities and sessions, and computes the volume-weighted mean paired difference; we report the 2.5th and 97.5th percentiles.

Table 9: Paired PSNR differences (dB), C16 minus comparator. Positive values favor ours. Intervals resample subjects rather than individual modalities.

| Dataset | Comparator | Δ PSNR (dB) | 95% interval |
|---|---|---|---|
| BraTS | Ours (C8) | +3.20 | [+3.17, +3.23] |
| BraTS | HunyuanVideo | +1.13 | [+1.10, +1.16] |
| BraTS | Wan | +1.81 | [+1.78, +1.85] |
| BraTS | CONFLUX | +9.94 | [+9.70, +10.17] |
| OASIS | Ours (C8) | +2.51 | [+2.49, +2.54] |
| OASIS | HunyuanVideo | +0.75 | [+0.74, +0.75] |
| OASIS | Wan | +1.26 | [+1.25, +1.27] |
| OASIS | CONFLUX | +14.48 | [+14.35, +14.62] |
| IXI | Ours (C8) | +3.34 | [+3.30, +3.37] |
| IXI | HunyuanVideo | +1.03 | [+0.99, +1.06] |
| IXI | Wan | +1.51 | [+1.48, +1.54] |
| IXI | CONFLUX | +7.71 | [+7.22, +8.23] |
| MSD Brain | Ours (C8) | +3.02 | [+2.98, +3.06] |
| MSD Brain | HunyuanVideo | +1.07 | [+1.04, +1.10] |
| MSD Brain | Wan | +1.55 | [+1.52, +1.58] |
| MSD Brain | CONFLUX | +9.30 | [+8.85, +9.74] |

## F Input Shape Coverage

Shapes describe reference volumes after preprocessing and before model-specific padding or tiling. Each axis range gives its marginal minimum and maximum; the three extrema may occur in different volumes. Aspect ratio is the longest axis divided by the shortest.

Table 10: Reference-volume shape coverage. The three axis ranges are separated by slashes.

| Dataset | $n$ | Distinct shapes | Axis ranges | Max. aspect |
|---|---|---|---|---|
| BraTS | 1280 | 1 | 155–155/240–240/240–240 | 1.55 |
| OASIS | 562 | 1 | 160–160/256–256/256–256 | 1.60 |
| IXI | 200 | 4 | 144–180/240–240/240–240 | 1.67 |
| MSD Brain | 199 | 1 | 155–155/240–240/240–240 | 1.55 |
| BrainTumor-256K | 400 | 1 | 189–189/233–233/197–197 | 1.23 |
| AMOS MRI | 101 | 36 | 175–500/120–400/346–500 | 3.75 |
| Knee MRI | 320 | 21 | 84–150/140–140/140–140 | 1.67 |
| Heart MRI | 30 | 11 | 123–192/300–400/300–400 | 3.25 |
| Prostate MRI | 96 | 9 | 44–86/192–232/192–232 | 5.27 |
| Liver CT | 200 | 196 | 74–1800/237–512/285–512 | 6.92 |
| Lung CT | 95 | 95 | 169–636/306–500/306–500 | 2.37 |
| Pancreas CT | 200 | 196 | 120–670/277–500/277–500 | 3.22 |
| KiTS19 CT | 200 | 200 | 145–755/224–533/224–533 | 3.45 |

Table 11: Autoencoding fidelity on the internal BrainTumor-256K validation cohort. Rows are ordered by PSNR; bold marks the best mean in each column.

| Model | PSNR ↑ | MS-SSIM ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| Ours (C16) | **29.09** | **0.9894** | 0.9174 | 0.1217 |
| CogVideoX | 28.91 | 0.9879 | 0.9153 | 0.1762 |
| HunyuanVideo | 28.78 | 0.9888 | **0.9176** | 0.1193 |
| Wan | 28.04 | 0.9867 | 0.9056 | **0.1157** |
| MAISI-MR | 26.81 | 0.9601 | 0.7510 | 0.1373 |
| Ours (C8) | 26.56 | 0.9800 | 0.8719 | 0.1484 |
| MedVAE-f4c1 | 26.16 | 0.9763 | 0.8553 | 0.1558 |
| MAISI-CT | 26.16 | 0.9534 | 0.7357 | 0.1699 |
| Patch-Volume | 22.03 | 0.9283 | 0.7427 | 0.2807 |
| MedVAE-f8c1 | 21.91 | 0.9091 | 0.7073 | 0.3331 |
| Brain-LDM | 21.10 | 0.9019 | 0.6686 | 0.4040 |
| CONFLUX | 19.93 | 0.8658 | 0.5102 | 0.4381 |

## G Multi-metric Autoencoding and Coverage

This appendix reports PSNR, MS-SSIM, SSIM and LPIPS for all evaluated models, including both C16 and C8. Within each dataset, comparisons use the intersection of volumes with finite values for all four metrics in every model; Table 12 gives the resulting sample counts and Table 13 reports our available-case means on each full input pool. CT comparisons use volumes completed by all models.

Table 12: Evaluation coverage: volumes with finite values for all four metrics. Shared denotes the intersection across models.

| Dataset | Total | C16 | C8 | Hunyuan | Wan | Cog | Med-f4 | Med-f8 | PV | MAISI-MR | MAISI-CT | Brain-LDM | CONFLUX | Shared |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BraTS | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 | 1280 |
| OASIS | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 | 562 |
| IXI | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |
| MSD Brain | 200 | 199 | 199 | 199 | 199 | 199 | 199 | 199 | 199 | 199 | 199 | 199 | 199 | 199 |
| BrainTumor-256K | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 | 400 |
| AMOS MRI | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 | 101 |
| Knee MRI | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 | 320 |
| Heart MRI | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |
| Prostate MRI | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 |
| Liver CT | 200 | 200 | 200 | 200 | 200 | 200 | 199 | 199 | 136 | 195 | 195 | 175 | 200 | 131 |
| Lung CT | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 92 | 95 | 95 | 95 | 95 | 92 |
| Pancreas CT | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 195 | 200 | 200 | 199 | 200 | 195 |
| KiTS19 CT | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 153 | 198 | 198 | 181 | 200 | 153 |

Table 13: Available-case means for C16 and C8 on each full input pool. $n$ counts volumes with all four metrics valid.

| Dataset | Variant | $n$ | PSNR | MS-SSIM | SSIM | LPIPS |
|---|---|---|---|---|---|---|
| BraTS | Ours (C16) | 1280 | 31.79 | 0.9951 | 0.9565 | 0.0571 |
| BraTS | Ours (C8) | 1280 | 28.59 | 0.9884 | 0.9211 | 0.0901 |
| OASIS | Ours (C16) | 562 | 29.46 | 0.9951 | 0.9492 | 0.0712 |
| OASIS | Ours (C8) | 562 | 26.95 | 0.9894 | 0.9161 | 0.0929 |
| IXI | Ours (C16) | 200 | 31.81 | 0.9969 | 0.9710 | 0.0531 |
| IXI | Ours (C8) | 200 | 28.47 | 0.9917 | 0.9410 | 0.0822 |
| MSD Brain | Ours (C16) | 199 | 30.91 | 0.9951 | 0.9558 | 0.0625 |
| MSD Brain | Ours (C8) | 199 | 27.88 | 0.9885 | 0.9189 | 0.0986 |
| BrainTumor-256K | Ours (C16) | 400 | 29.09 | 0.9894 | 0.9174 | 0.1217 |
| BrainTumor-256K | Ours (C8) | 400 | 26.56 | 0.9800 | 0.8719 | 0.1484 |
| AMOS MRI | Ours (C16) | 101 | 31.57 | 0.9790 | 0.8797 | 0.1513 |
| AMOS MRI | Ours (C8) | 101 | 27.91 | 0.9423 | 0.7235 | 0.2698 |
| Knee MRI | Ours (C16) | 320 | 28.07 | 0.9720 | 0.7764 | 0.2239 |
| Knee MRI | Ours (C8) | 320 | 26.15 | 0.9550 | 0.7166 | 0.3185 |
| Heart MRI | Ours (C16) | 30 | 32.34 | 0.9806 | 0.8901 | 0.1553 |
| Heart MRI | Ours (C8) | 30 | 28.69 | 0.9543 | 0.7801 | 0.2884 |
| Prostate MRI | Ours (C16) | 96 | 27.66 | 0.9788 | 0.8383 | 0.1739 |
| Prostate MRI | Ours (C8) | 96 | 24.45 | 0.9543 | 0.7225 | 0.3076 |
| Liver CT | Ours (C16) | 200 | 33.56 | 0.9718 | 0.8449 | 0.1899 |
| Liver CT | Ours (C8) | 200 | 29.76 | 0.9498 | 0.7574 | 0.2941 |
| Lung CT | Ours (C16) | 95 | 27.85 | 0.9361 | 0.6327 | 0.2983 |
| Lung CT | Ours (C8) | 95 | 25.34 | 0.9045 | 0.5262 | 0.4045 |
| Pancreas CT | Ours (C16) | 200 | 33.50 | 0.9667 | 0.8347 | 0.1996 |
| Pancreas CT | Ours (C8) | 200 | 29.42 | 0.9406 | 0.7229 | 0.3065 |
| KiTS19 CT | Ours (C16) | 200 | 34.53 | 0.9799 | 0.8828 | 0.1836 |
| KiTS19 CT | Ours (C8) | 200 | 30.54 | 0.9614 | 0.8097 | 0.2881 |

Table 14: Matched-volume reconstruction metrics for BraTS and OASIS. All models use the same valid volumes within each dataset. Rows are ordered by PSNR within each dataset; bold marks the best mean for each metric.

**BraTS (1280 matched volumes)**

| Model | PSNR ↑ | MS-SSIM ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| Ours (C16) | **31.79** | **0.9951** | **0.9565** | **0.0571** |
| HunyuanVideo | 30.66 | 0.9921 | 0.9408 | 0.0776 |
| Wan | 29.98 | 0.9901 | 0.9205 | 0.0881 |
| Ours (C8) | 28.59 | 0.9884 | 0.9211 | 0.0901 |
| MedVAE-f4c1 | 28.07 | 0.9843 | 0.9008 | 0.1223 |
| MAISI-MR | 27.68 | 0.9447 | 0.5413 | 0.1210 |
| MAISI-CT | 26.12 | 0.9350 | 0.5366 | 0.1620 |
| MedVAE-f8c1 | 22.88 | 0.9141 | 0.7338 | 0.3514 |
| Patch-Volume | 22.85 | 0.9164 | 0.7555 | 0.2339 |
| CONFLUX | 21.86 | 0.9655 | 0.8661 | 0.3018 |
| CogVideoX | 20.95 | 0.9192 | 0.7565 | 0.1639 |
| Brain-LDM | 18.61 | 0.7996 | 0.3318 | 0.6564 |

**OASIS (562 matched volumes)**

| Model | PSNR ↑ | MS-SSIM ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| Ours (C16) | **29.46** | **0.9951** | **0.9492** | **0.0712** |
| HunyuanVideo | 28.72 | 0.9927 | 0.9383 | 0.0895 |
| CogVideoX | 28.71 | 0.9904 | 0.9229 | 0.1455 |
| Wan | 28.20 | 0.9905 | 0.9190 | 0.0891 |
| Ours (C8) | 26.95 | 0.9894 | 0.9161 | 0.0929 |
| MAISI-MR | 26.12 | 0.9404 | 0.5408 | 0.1340 |
| MedVAE-f4c1 | 25.67 | 0.9831 | 0.8770 | 0.1297 |
| MAISI-CT | 25.39 | 0.9352 | 0.5480 | 0.1887 |
| Patch-Volume | 22.01 | 0.9514 | 0.8054 | 0.2193 |
| MedVAE-f8c1 | 20.23 | 0.8991 | 0.6742 | 0.3698 |
| Brain-LDM | 17.88 | 0.8272 | 0.3397 | 0.6257 |
| CONFLUX | 14.98 | 0.8368 | 0.4710 | 0.6021 |

Table 15: Matched-volume reconstruction metrics for IXI and MSD Brain. All models use the same valid volumes within each dataset. Rows are ordered by PSNR within each dataset; bold marks the best mean for each metric.

**IXI (200 matched volumes)**

| Model | PSNR ↑ | MS-SSIM ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| Ours (C16) | **31.81** | **0.9969** | **0.9710** | **0.0531** |
| HunyuanVideo | 30.79 | 0.9948 | 0.9611 | 0.0672 |
| Wan | 30.30 | 0.9929 | 0.9439 | 0.0691 |
| Ours (C8) | 28.47 | 0.9917 | 0.9410 | 0.0822 |
| MedVAE-f4c1 | 27.85 | 0.9879 | 0.9238 | 0.1300 |
| MAISI-MR | 27.75 | 0.9405 | 0.5499 | 0.1105 |
| MAISI-CT | 25.54 | 0.9289 | 0.5465 | 0.1688 |
| CogVideoX | 24.46 | 0.9656 | 0.8549 | 0.1233 |
| CONFLUX | 24.10 | 0.9681 | 0.8737 | 0.2570 |
| Patch-Volume | 22.80 | 0.9455 | 0.8043 | 0.2109 |
| MedVAE-f8c1 | 21.88 | 0.9076 | 0.7239 | 0.3520 |
| Brain-LDM | 18.20 | 0.8030 | 0.3179 | 0.6818 |

**MSD Brain (199 matched volumes)**

| Model | PSNR ↑ | MS-SSIM ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| Ours (C16) | **30.91** | **0.9951** | **0.9558** | **0.0625** |
| HunyuanVideo | 29.84 | 0.9923 | 0.9400 | 0.0845 |
| Wan | 29.36 | 0.9904 | 0.9194 | 0.1000 |
| Ours (C8) | 27.88 | 0.9885 | 0.9189 | 0.0986 |
| MAISI-MR | 27.61 | 0.9478 | 0.5540 | 0.1234 |
| MedVAE-f4c1 | 27.56 | 0.9847 | 0.8999 | 0.1293 |
| MAISI-CT | 26.21 | 0.9389 | 0.5546 | 0.1598 |
| MedVAE-f8c1 | 22.94 | 0.9199 | 0.7387 | 0.3929 |
| Patch-Volume | 22.74 | 0.9114 | 0.7416 | 0.2741 |
| CONFLUX | 21.60 | 0.9661 | 0.8627 | 0.3521 |
| CogVideoX | 20.60 | 0.9217 | 0.7597 | 0.1739 |
| Brain-LDM | 18.55 | 0.7944 | 0.3150 | 0.7053 |

Table 16: PSNR (dB) rankings on non-brain MRI and CT. Higher is better. Each dataset column is sorted independently using unrounded matched-volume means. Entries give model name (score); the best entry is bold. Solid and dashed boxes denote C16 and C8, respectively. $n$ is the shared valid-volume count.

| Rank | AMOS MRI $n = 101$ | Knee MRI $n = 320$ | Heart MRI $n = 30$ | Prostate MRI $n = 96$ | Liver CT $n = 131$ | Lung CT $n = 92$ | Pancreas CT $n = 195$ | KiTS19 CT $n = 153$ |
|---|---|---|---|---|---|---|---|---|
| 1 | **HunyuanVideo (33.23)** | **HunyuanVideo (28.46)** | **HunyuanVideo (33.20)** | **HunyuanVideo (28.52)** | **HunyuanVideo (36.65)** | **HunyuanVideo (29.21)** | **HunyuanVideo (37.32)** | **HunyuanVideo (37.86)** |
| 2 | Wan (31.95) | Ours (C16) (28.07) | Wan (32.41) | Wan (27.75) | Wan (35.54) | Wan (28.60) | Wan (36.30) | Wan (36.33) |
| 3 | Ours (C16) (31.57) | Wan (27.77) | Ours (C16) (32.34) | Ours (C16) (27.66) | Ours (C16) (33.64) | Ours (C16) (27.86) | Ours (C16) (33.48) | Ours (C16) (34.37) |
| 4 | CogVideoX (31.24) | Ours (C8) (26.15) | MedVAE-f4c1 (30.11) | CogVideoX (27.23) | CogVideoX (31.88) | MedVAE-f4c1 (27.20) | CogVideoX (30.47) | CogVideoX (31.47) |
| 5 | MedVAE-f4c1 (29.29) | MAISI-MR (25.87) | MAISI-MR (30.05) | MAISI-MR (25.74) | MedVAE-f4c1 (30.54) | CogVideoX (26.72) | MedVAE-f4c1 (30.26) | Ours (C8) (30.53) |
| 6 | MAISI-MR (28.49) | MAISI-CT (25.44) | CogVideoX (29.41) | MAISI-CT (25.31) | Ours (C8) (29.99) | CONFLUX (25.95) | Ours (C8) (29.41) | MedVAE-f4c1 (29.73) |
| 7 | Ours (C8) (27.91) | CogVideoX (25.02) | Ours (C8) (28.69) | Ours (C8) (24.45) | Patch-Volume (29.14) | Ours (C8) (25.35) | Patch-Volume (28.16) | Patch-Volume (28.38) |
| 8 | MAISI-CT (26.83) | MedVAE-f4c1 (24.98) | MAISI-CT (28.55) | MedVAE-f4c1 (24.07) | CONFLUX (28.42) | Patch-Volume (24.75) | CONFLUX (26.42) | MAISI-MR (28.25) |
| 9 | Patch-Volume (25.19) | Patch-Volume (22.95) | Patch-Volume (25.77) | Patch-Volume (21.14) | MAISI-MR (27.56) | MAISI-MR (24.72) | MAISI-CT (23.40) | MAISI-CT (28.18) |
| 10 | MedVAE-f8c1 (20.49) | MedVAE-f8c1 (22.52) | Brain-LDM (21.88) | MedVAE-f8c1 (20.19) | MAISI-CT (27.42) | MAISI-CT (23.04) | MAISI-MR (22.76) | CONFLUX (25.31) |
| 11 | CONFLUX (19.12) | Brain-LDM (17.61) | MedVAE-f8c1 (19.67) | Brain-LDM (17.24) | MedVAE-f8c1 (20.65) | MedVAE-f8c1 (17.99) | MedVAE-f8c1 (19.96) | MedVAE-f8c1 (19.74) |
| 12 | Brain-LDM (17.32) | CONFLUX (14.06) | CONFLUX (19.61) | CONFLUX (16.95) | Brain-LDM (7.98) | Brain-LDM (9.64) | Brain-LDM (9.16) | Brain-LDM (7.21) |

Table 17: MS-SSIM rankings on non-brain MRI and CT. Higher is better. Each dataset column is sorted independently using unrounded matched-volume means. Entries give model name (score); the best entry is bold. Solid and dashed boxes denote C16 and C8, respectively. $n$ is the shared valid-volume count.

| Rank | AMOS MRI $n = 101$ | Knee MRI $n = 320$ | Heart MRI $n = 30$ | Prostate MRI $n = 96$ | Liver CT $n = 131$ | Lung CT $n = 92$ | Pancreas CT $n = 195$ | KiTS19 CT $n = 153$ |
|---|---|---|---|---|---|---|---|---|
| 1 | **HunyuanVideo (0.9928)** | **HunyuanVideo (0.9739)** | **HunyuanVideo (0.9905)** | **HunyuanVideo (0.9825)** | **HunyuanVideo (0.9908)** | **HunyuanVideo (0.9688)** | **HunyuanVideo (0.9933)** | **HunyuanVideo (0.9924)** |
| 2 | Wan (0.9893) | Ours (C16) (0.9720) | Wan (0.9885) | Wan (0.9798) | Wan (0.9889) | Wan (0.9642) | Wan (0.9915) | Wan (0.9909) |
| 3 | CogVideoX (0.9841) | Wan (0.9709) | Ours (C16) (0.9806) | Ours (C16) (0.9788) | CogVideoX (0.9806) | MedVAE-f4c1 (0.9569) | CogVideoX (0.9840) | CogVideoX (0.9831) |
| 4 | MedVAE-f4c1 (0.9811) | MAISI-MR (0.9607) | MedVAE-f4c1 (0.9775) | CogVideoX (0.9770) | MedVAE-f4c1 (0.9755) | CogVideoX (0.9541) | MedVAE-f4c1 (0.9828) | Ours (C16) (0.9807) |
| 5 | Ours (C16) (0.9790) | MAISI-CT (0.9566) | CogVideoX (0.9722) | MedVAE-f4c1 (0.9580) | Ours (C16) (0.9736) | Ours (C16) (0.9361) | Ours (C16) (0.9666) | MedVAE-f4c1 (0.9779) |
| 6 | Ours (C8) (0.9423) | Ours (C8) (0.9550) | MAISI-MR (0.9644) | Ours (C8) (0.9543) | MAISI-MR (0.9604) | MAISI-MR (0.9229) | Patch-Volume (0.9573) | MAISI-MR (0.9683) |
| 7 | Patch-Volume (0.9405) | MedVAE-f4c1 (0.9533) | MAISI-CT (0.9557) | MAISI-MR (0.9543) | MAISI-CT (0.9597) | MAISI-CT (0.9155) | Ours (C8) (0.9407) | MAISI-CT (0.9675) |
| 8 | MAISI-MR (0.9329) | CogVideoX (0.9446) | Ours (C8) (0.9543) | MAISI-CT (0.9502) | Patch-Volume (0.9547) | Patch-Volume (0.9150) | MAISI-CT (0.9370) | Ours (C8) (0.9640) |
| 9 | MAISI-CT (0.9229) | Patch-Volume (0.9003) | Patch-Volume (0.9227) | Patch-Volume (0.8916) | Ours (C8) (0.9541) | CONFLUX (0.9089) | MAISI-MR (0.9352) | Patch-Volume (0.9508) |
| 10 | CONFLUX (0.8617) | MedVAE-f8c1 (0.8794) | Brain-LDM (0.8518) | MedVAE-f8c1 (0.8453) | CONFLUX (0.9376) | Ours (C8) (0.9046) | CONFLUX (0.9185) | CONFLUX (0.9360) |
| 11 | MedVAE-f8c1 (0.8133) | Brain-LDM (0.8371) | CONFLUX (0.8476) | CONFLUX (0.8069) | MedVAE-f8c1 (0.8360) | MedVAE-f8c1 (0.7633) | MedVAE-f8c1 (0.8355) | MedVAE-f8c1 (0.8281) |
| 12 | Brain-LDM (0.6681) | CONFLUX (0.7413) | MedVAE-f8c1 (0.7290) | Brain-LDM (0.7771) | Brain-LDM (0.3824) | Brain-LDM (0.3815) | Brain-LDM (0.4577) | Brain-LDM (0.4063) |

Table 18: SSIM rankings on non-brain MRI and CT. Higher is better. Each dataset column is sorted independently using unrounded matched-volume means. Entries give model name (score); the best entry is bold. Solid and dashed boxes denote C16 and C8, respectively. $n$ is the shared valid-volume count.

| Rank | AMOS MRI $n = 101$ | Knee MRI $n = 320$ | Heart MRI $n = 30$ | Prostate MRI $n = 96$ | Liver CT $n = 131$ | Lung CT $n = 92$ | Pancreas CT $n = 195$ | KiTS19 CT $n = 153$ |
|---|---|---|---|---|---|---|---|---|
| 1 | **HunyuanVideo (0.9471)** | **HunyuanVideo (0.7849)** | **HunyuanVideo (0.9356)** | **HunyuanVideo (0.8620)** | **HunyuanVideo (0.9244)** | **HunyuanVideo (0.7637)** | **HunyuanVideo (0.9466)** | **HunyuanVideo (0.9396)** |
| 2 | Wan (0.9271) | Ours (C16) (0.7764) | Wan (0.9220) | Wan (0.8426) | Wan (0.9124) | Wan (0.7402) | Wan (0.9338) | Wan (0.9278) |
| 3 | CogVideoX (0.9165) | Wan (0.7604) | Ours (C16) (0.8901) | Ours (C16) (0.8383) | CogVideoX (0.8963) | CogVideoX (0.7307) | CogVideoX (0.9173) | CogVideoX (0.9113) |
| 4 | MedVAE-f4c1 (0.8868) | Ours (C8) (0.7166) | CogVideoX (0.8863) | CogVideoX (0.8352) | MedVAE-f4c1 (0.8714) | MedVAE-f4c1 (0.7118) | MedVAE-f4c1 (0.9007) | MedVAE-f4c1 (0.8884) |
| 5 | Ours (C16) (0.8797) | CogVideoX (0.7121) | MedVAE-f4c1 (0.8758) | MedVAE-f4c1 (0.7359) | Ours (C16) (0.8520) | Ours (C16) (0.6337) | Ours (C16) (0.8344) | Ours (C16) (0.8852) |
| 6 | Patch-Volume (0.7563) | MAISI-MR (0.7101) | MAISI-MR (0.8084) | Ours (C8) (0.7225) | Patch-Volume (0.8046) | Patch-Volume (0.6139) | Patch-Volume (0.8301) | MAISI-CT (0.8279) |
| 7 | Ours (C8) (0.7235) | MAISI-CT (0.7097) | MAISI-CT (0.7920) | MAISI-MR (0.7160) | MAISI-CT (0.7982) | MAISI-MR (0.5682) | MAISI-CT (0.7417) | MAISI-MR (0.8245) |
| 8 | MAISI-MR (0.6922) | MedVAE-f4c1 (0.6799) | Ours (C8) (0.7801) | MAISI-CT (0.7149) | MAISI-MR (0.7957) | MAISI-CT (0.5674) | MAISI-MR (0.7381) | Ours (C8) (0.8187) |
| 9 | MAISI-CT (0.6705) | Patch-Volume (0.5613) | Patch-Volume (0.7069) | Patch-Volume (0.5540) | Ours (C8) (0.7728) | CONFLUX (0.5504) | Ours (C8) (0.7229) | Patch-Volume (0.8113) |
| 10 | MedVAE-f8c1 (0.6186) | MedVAE-f8c1 (0.5464) | CONFLUX (0.5567) | MedVAE-f8c1 (0.5124) | CONFLUX (0.7393) | Ours (C8) (0.5268) | MedVAE-f8c1 (0.6895) | CONFLUX (0.7675) |
| 11 | CONFLUX (0.6164) | Brain-LDM (0.3969) | Brain-LDM (0.5503) | CONFLUX (0.4472) | MedVAE-f8c1 (0.6868) | MedVAE-f8c1 (0.4922) | CONFLUX (0.6779) | MedVAE-f8c1 (0.6936) |
| 12 | Brain-LDM (0.4325) | CONFLUX (0.2524) | MedVAE-f8c1 (0.4817) | Brain-LDM (0.3233) | Brain-LDM (0.1611) | Brain-LDM (0.1357) | Brain-LDM (0.2196) | Brain-LDM (0.1699) |

Table 19: LPIPS rankings on non-brain MRI and CT. Lower is better. Each dataset column is sorted independently using unrounded matched-volume means. Entries give model name (score); the best entry is bold. Solid and dashed boxes denote C16 and C8, respectively. $n$ is the shared valid-volume count.

| Rank | AMOS MRI $n = 101$ | Knee MRI $n = 320$ | Heart MRI $n = 30$ | Prostate MRI $n = 96$ | Liver CT $n = 131$ | Lung CT $n = 92$ | Pancreas CT $n = 195$ | KiTS19 CT $n = 153$ |
|---|---|---|---|---|---|---|---|---|
| 1 | **HunyuanVideo (0.1001)** | **HunyuanVideo (0.1885)** | **HunyuanVideo (0.1318)** | **Ours (C16) (0.1739)** | **HunyuanVideo (0.1150)** | **Wan (0.1662)** | **HunyuanVideo (0.0851)** | **HunyuanVideo (0.1018)** |
| 2 | Wan (0.1242) | Wan (0.2161) | Wan (0.1489) | HunyuanVideo (0.1845) | Wan (0.1221) | HunyuanVideo (0.1713) | Wan (0.0936) | Wan (0.1128) |
| 3 | CogVideoX (0.1357) | Ours (C16) (0.2239) | Ours (C16) (0.1553) | Wan (0.2319) | CogVideoX (0.1669) | MedVAE-f4c1 (0.1973) | CogVideoX (0.1395) | CogVideoX (0.1524) |
| 4 | Ours (C16) (0.1513) | MAISI-MR (0.2698) | MAISI-MR (0.1620) | CogVideoX (0.2919) | MedVAE-f4c1 (0.1773) | MAISI-MR (0.2366) | MedVAE-f4c1 (0.1989) | Ours (C16) (0.1808) |
| 5 | MAISI-MR (0.1721) | MAISI-CT (0.3115) | CogVideoX (0.1697) | MAISI-MR (0.3010) | Ours (C16) (0.1858) | Patch-Volume (0.2389) | Ours (C16) (0.1996) | MAISI-MR (0.1862) |
| 6 | MAISI-CT (0.2208) | Ours (C8) (0.3185) | MAISI-CT (0.2111) | Ours (C8) (0.3076) | MAISI-MR (0.1894) | CogVideoX (0.2643) | Patch-Volume (0.2633) | MAISI-CT (0.2043) |
| 7 | MedVAE-f4c1 (0.2247) | MedVAE-f4c1 (0.3805) | MedVAE-f4c1 (0.2129) | MAISI-CT (0.3306) | MAISI-CT (0.2102) | Ours (C16) (0.2974) | MAISI-MR (0.2823) | MedVAE-f4c1 (0.2079) |
| 8 | Ours (C8) (0.2698) | CogVideoX (0.3882) | Ours (C8) (0.2884) | MedVAE-f4c1 (0.4374) | Patch-Volume (0.2193) | MAISI-CT (0.2983) | MAISI-CT (0.2889) | Ours (C8) (0.2798) |
| 9 | Patch-Volume (0.3494) | Patch-Volume (0.4312) | Patch-Volume (0.3409) | Patch-Volume (0.5479) | Ours (C8) (0.2808) | Ours (C8) (0.4031) | Ours (C8) (0.3060) | Patch-Volume (0.2852) |
| 10 | CONFLUX (0.4751) | MedVAE-f8c1 (0.6198) | CONFLUX (0.5189) | MedVAE-f8c1 (0.7991) | CONFLUX (0.4377) | MedVAE-f8c1 (0.5393) | CONFLUX (0.4362) | CONFLUX (0.4752) |
| 11 | MedVAE-f8c1 (0.5052) | Brain-LDM (0.8619) | Brain-LDM (0.5659) | CONFLUX (0.8060) | MedVAE-f8c1 (0.4498) | CONFLUX (0.6297) | MedVAE-f8c1 (0.4746) | MedVAE-f8c1 (0.5136) |
| 12 | Brain-LDM (0.8317) | CONFLUX (0.9516) | MedVAE-f8c1 (0.5853) | Brain-LDM (0.9673) | Brain-LDM (1.4832) | Brain-LDM (1.2653) | Brain-LDM (1.3933) | Brain-LDM (1.5630) |

Table 20: Patient-paired Dice differences in percentage points, ours minus comparator. Pointwise intervals use 2,000 patient-bootstrap replicates and no multiplicity adjustment. Models and segmenter are fixed across replicates.

| Comparator | Region | Patients | Difference | 95% interval |
|---|---|---|---|---|
| Original | WT | 165 | -0.24 | [-0.53, +0.04] |
| HunyuanVideo | WT | 165 | -1.49 | [-2.40, -0.47] |
| Original | TC | 165 | -0.46 | [-0.96, -0.08] |
| HunyuanVideo | TC | 165 | +1.58 | [+0.68, +2.62] |
| Original | ET | 158 | -0.73 | [-1.20, -0.36] |
| HunyuanVideo | ET | 158 | +1.57 | [+0.90, +2.32] |

## H TUMOR-REGION PRESERVATION PROTOCOL AND RESULTS

**Cohort and comparison.** We use 165 preoperative patients from a fixed BraTS-GLI validation partition of the BraTS benchmark (Menze et al., 2015), with one preoperative examination per patient. Each examination includes non-contrast T1, contrast-enhanced T1, T2 and FLAIR images with a tumor annotation. Our brain VAE and the released HunyuanVideo VAE reconstruct each modality independently using posterior means; the frozen segmenter then receives the four aligned reconstructed modalities.

**Shared preprocessing.** Inputs are already registered and skull-stripped on a 1 mm grid. N4 correction (Tustison et al., 2010) is applied once to the original modalities before branching. Each original modality determines its 0.5th and 99.5th intensity percentiles; the clipped interval is mapped to $[0, 1]$ for reconstruction and the same constants restore the intensity scale for every input condition. The original-input condition uses this same corrected and clipped reference. NS-HGlio receives these prepared inputs with its own registration, skull stripping and bias correction disabled. Annotations are withheld from both reconstruction and segmentation.

**Region definitions and scoring.** Reference labels 1, 2 and 3 represent non-enhancing tumor core, surrounding non-enhancing FLAIR hyperintensity, and enhancing tumor; predicted labels 1, 2 and 3 represent necrosis, enhancing tumor and edema. Reference WT/TC/ET sets are therefore $\{1, 2, 3\}/\{1, 3\}/\{3\}$, while predicted sets are $\{1, 2, 3\}/\{1, 2\}/\{2\}$. Dice is computed per region; an empty prediction for a present region scores zero, and absent reference regions are excluded. WT and TC have 165 positive-reference patients each; ET has 158, leaving seven ET-absent cases. Table 20 reports patient-paired Dice differences of our reconstructions relative to the original inputs and to HunyuanVideo reconstructions.

## I DIT RESOURCE BENCHMARK PROTOCOL

We isolate the computational effect of the latent interface by profiling complete DiT forward, backward and AdamW optimizer steps on synthetic tensors corresponding to a $192 \times 192 \times 144$ image. The latent patch size is two unless stated otherwise. Each run uses one H200 GPU, batch size one, BF16 autocast with FP32 parameters, SDPA, five warmup steps and twenty measured steps. Compilation and activation checkpointing are disabled. VAE execution and data loading are outside the timing boundary. Table 21 reports median step time, within-run standard deviation, and peak allocated and reserved memory.

The benchmark includes 633M, 988M and 2.01B DiT backbones. Within each size group, interface comparisons fix depth, width, attention-head count and optimizer settings; only the channel-dependent input/output projections differ.

## J CONDITIONAL GENERATION PROTOCOL

We use a 2B-parameter multimodal Diffusion Transformer of width 1728 with a frozen C16 VAE. A $192 \times 192 \times 144$ image maps to a $16 \times 24 \times 24 \times 18$ latent tensor. Both generation tasks use the same generator and VAE.

Table 21: DiT training resources for synthetic latent tensors. Interfaces are named by spatial reduction and channel count, with the corresponding VAEs in parentheses; p denotes the DiT latent patch size (two unless stated). SD is the standard deviation across measured steps within one run.

| DiT | Interface | Tokens | Step ms | SD ms | Alloc. GiB | Res. GiB |
|---|---|---|---|---|---|---|
| 633M | f8/C16 (ours, CONFLUX) | 1,296 | 64.8 | 0.3 | 9.85 | 10.93 |
| 633M | (4,8,8)/C16 (video VAEs) | 2,592 | 70.9 | 5.8 | 11.31 | 12.40 |
| 633M | f4/C4 (MAISI), p2 | 10,368 | 946.2 | 79.0 | 19.70 | 20.66 |
| 988M | f8/C16 (ours, CONFLUX) | 1,296 | 67.4 | 0.2 | 14.85 | 18.12 |
| 988M | (4,8,8)/C16 (video VAEs) | 2,592 | 78.9 | 0.1 | 16.55 | 18.99 |
| 2.01B | f8/C16 (ours, CONFLUX) | 1,296 | 113.2 | 0.2 | 30.23 | 36.26 |
| 2.01B | (4,8,8)/C16 (video VAEs) | 2,592 | 161.6 | 15.6 | 34.30 | 38.98 |
| 2.01B | f4/C4 (MAISI), p2 | 10,368 | 1828.5 | 343.1 | 57.92 | 60.74 |
| 2.01B | f4/C4 (MAISI), p4 | 1,296 | 118.3 | 26.4 | 30.24 | 36.25 |

Modality-controlled generation disables all reference-image inputs and conditions on the requested sequence. Mask-guided generation adds a tumor-label map through the reference-conditioning pathway. Sampling uses 50 Euler steps with a power time schedule of exponent three. Text guidance is five for modality-controlled synthesis; mask-guided synthesis uses text and reference guidance of four and six. Guidance settings and generator weights were selected on validation data before evaluation.

**Original-image references.** We evaluate the released generators and both tasks of our model against internal and BraTS cohorts of original MRI, using 400 generated and 400 reference volumes per supported sequence for each reported score. The internal cohort comprises four-sequence examinations drawn from the validation split, excluding examinations used for mask conditioning; the external cohort comprises BraTS subjects. Generated volumes and BraTS references are reoriented, resampled to 1 mm spacing, percentile-normalized and center-cropped or padded to $192 \times 192 \times 144$; internal references use the preprocessed 1 mm images on the same grid. Neither reference cohort is passed through a VAE. Slice-FID pools Inception-v3 features from axial slices between 20% and 85% of volume depth, with per-slice min–max normalization, and compares feature means and covariances (Heusel et al., 2017). Because the two reference cohorts differ in acquisition and preprocessing, absolute scores are not comparable across cohorts; we therefore compare models only within each reference cohort.

**Generation tasks and conditions.** NV-Generate-MR-Brain (NVIDIA, 2026), MONAI BraTS-LDM (MONAI Consortium, n.d.b) and WDM (Friedrich et al., 2024) use their released sampling pipelines for the supported sequences. MONAI BraTS-LDM denotes the BraTS FLAIR bundle; it is distinct from the UK Biobank Brain-LDM autoencoder used in reconstruction comparisons. Our mask-guided model and Med-DDPM (Dorjsembe et al., 2024) use the same 400 validation tumor masks, each supplying the condition for all four sequences. Our model generates the four sequences independently, conditioned on the tumor mask alone. Med-DDPM additionally receives the non-tumor brain-region label required by its conditioning interface. Masks are transformed into each model's input coordinates.

**Tumor-mask adherence.** We apply the frozen NS-HGlio segmenter (Abayazeed et al., 2023) to the four generated sequences for each of the 400 shared conditioning masks and compare the predicted whole-tumor region with the conditioning tumor region in a common coordinate frame. Table 22 reports case-averaged Dice, precision, recall and the predicted-to-conditioning tumor-volume ratio.

This protocol is conservative for our model. NS-HGlio is designed for four co-registered sequences from the same subject, whereas our model generates each sequence independently from the tumor mask alone, so the tumor is placed consistently but the surrounding anatomy differs across the four inputs. Med-DDPM additionally conditions on a shared brain-region label, which aligns non-lesional anatomy across sequences. Under this mismatch, our generated tumors are still recovered with a Dice of 0.82 and a higher recall than Med-DDPM; the lower precision and a volume ratio above one indicate that the segmenter extends predicted tumor regions into anatomically inconsistent surroundings rather than missing the conditioned lesion.

Table 22: Tumor-mask adherence measured by a frozen segmenter on generated MRI. Volume ratio is predicted tumor volume divided by conditioning tumor volume; the ideal value is one.

| Model | Dice ↑ | Precision ↑ | Recall ↑ | Volume ratio |
|---|---|---|---|---|
| Ours | 0.824 | 0.766 | 0.911 | 1.275 |
| Med-DDPM | 0.847 | 0.844 | 0.867 | 1.052 |

**Qualitative examples.** Figure 3 illustrates the supported generation tasks. The two paired mask-guided examples are selected from the shared conditioning cohort near the first and third quartiles of tumor volume, independently of generated appearance. Within each example, both models use the same mask and slice across all four sequences. Images share a canonical axial orientation and a 192 mm in-plane field of view, with volume-wise percentile normalization. Sequence-controlled slices use the midpoint of the occupied superior–inferior extent; mask-guided slices use the largest tumor cross-section for each lesion. All outputs are resampled to a common orientation and spacing before display.

## K Additional Mask-guided Generation Examples

Figures 6 and 7 compare five cases selected at the 10th, 30th, 50th, 70th and 90th percentiles of conditioning tumor volume, independently of generated appearance. Each comparison uses the same mask and axial slice across both models and all four sequences. Slices show the largest conditioning-tumor cross-section. All panels use a shared spatial field of view and display range after the volume-wise normalization described in Appendix J.

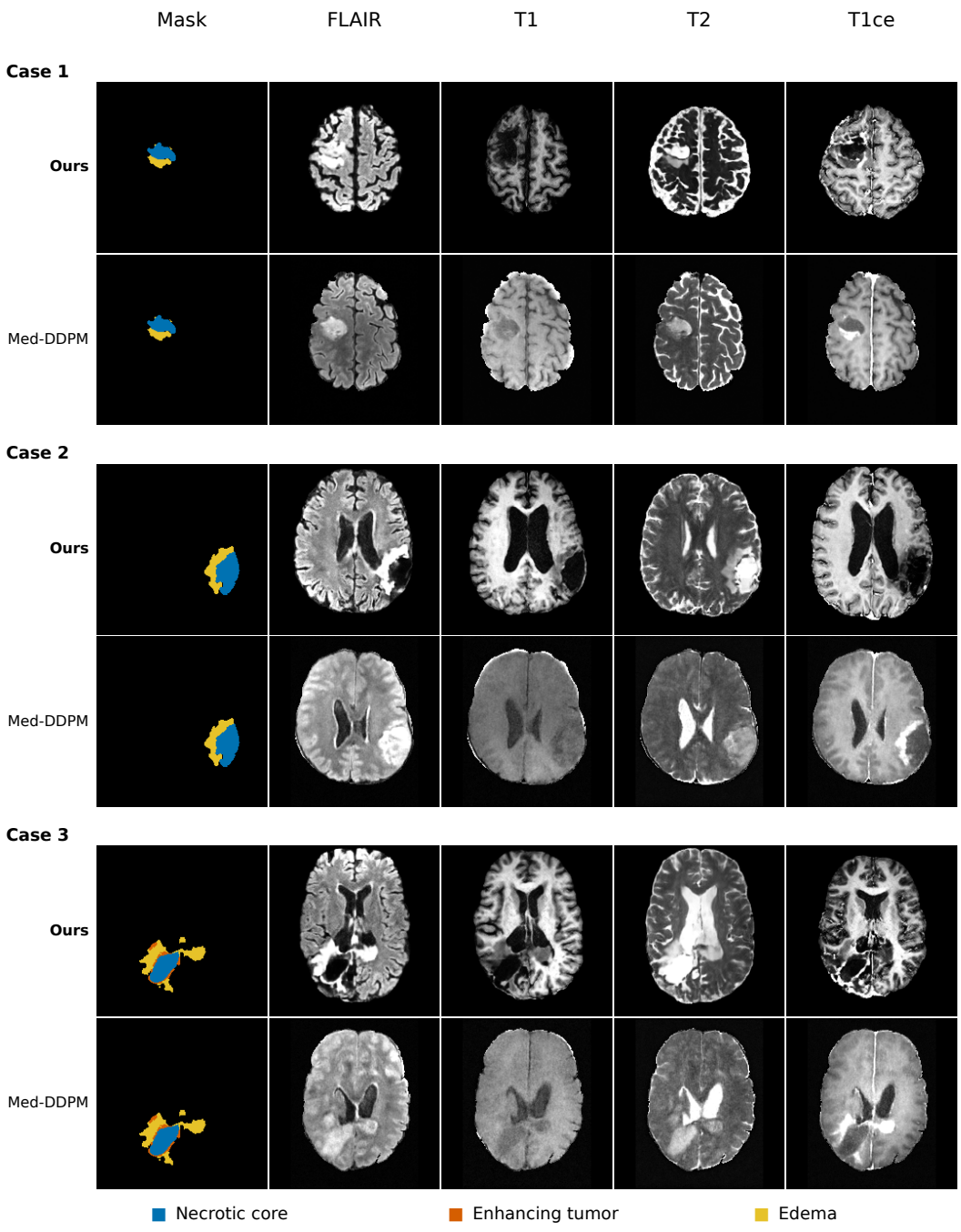


Figure 6: Mask-guided generation under shared tumor conditions, cases 1–3. Each pair shows Ours above Med-DDPM across four MRI sequences.

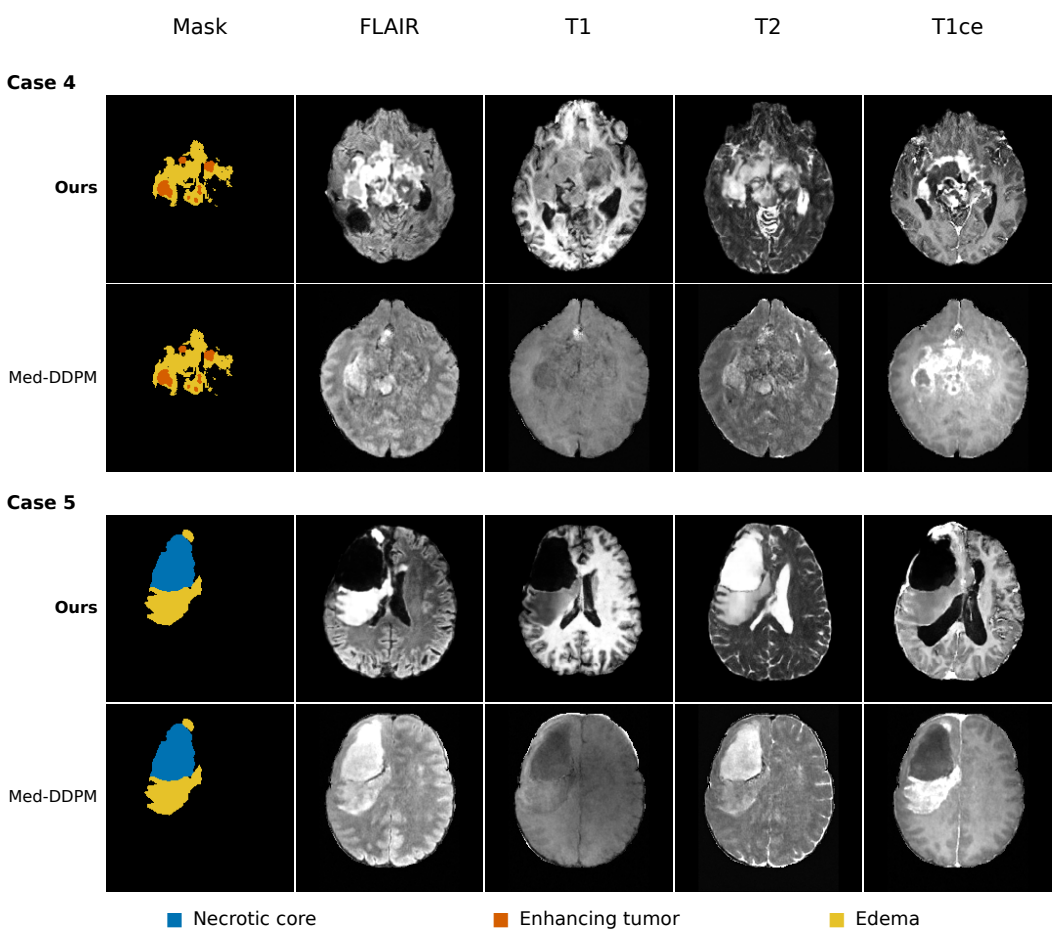


Figure 7: Mask-guided generation under shared tumor conditions, cases 4–5. Layout and display settings match Figure 6.

## L Additional Reconstruction Examples

Figures 8 and 9 show C16 and C8 reconstructions of tumor-centered regions across four MRI sequences. Cases span a range of regional reconstruction quality, illustrating how tumor boundaries and surrounding anatomy are preserved under $8 \times 8 \times 8$ compression.

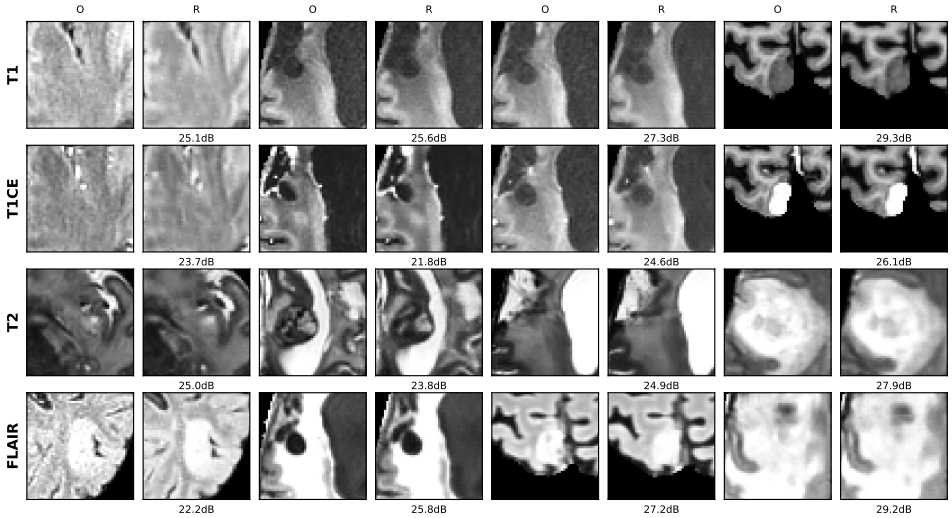


Figure 8: C16 original–reconstruction pairs spanning a range of tumor-region reconstruction quality. O and R denote original and reconstruction; rows show four MRI sequences, and annotations report PSNR (dB) in tumor-centered $48^3$ regions. Cases are ordered by regional MS-SSIM within the selected $[0.92, 1.0]$ band.

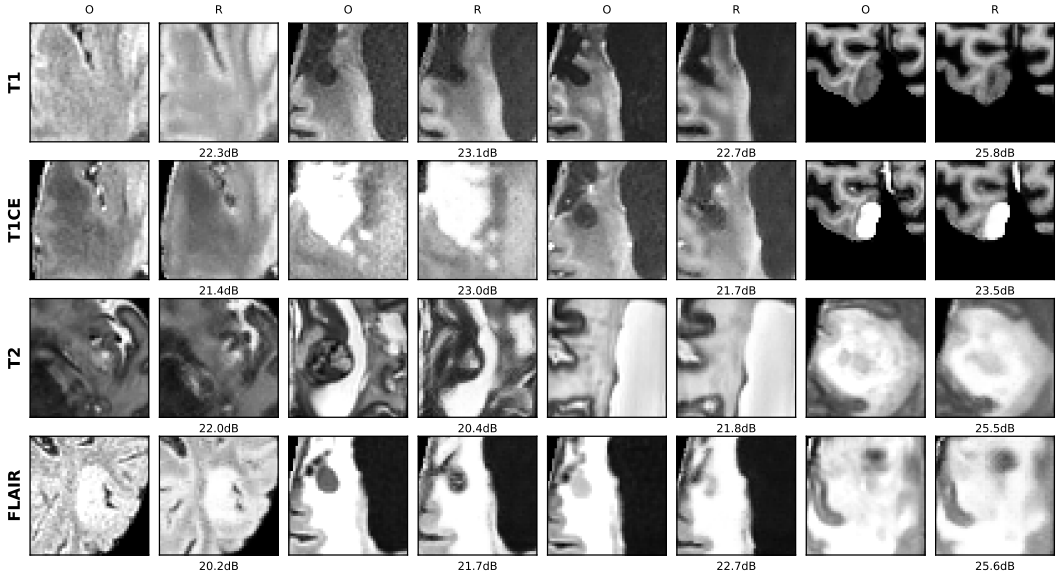


Figure 9: C8 original–reconstruction pairs across four MRI sequences, with O and R as in Figure 8. Annotations report regional PSNR (dB) in tumor-centered $48^3$ regions. Examples span the selected regional MS-SSIM band.