MEDGEN3D: UNIFIED 3D MEDICAL PREDICTION WITH A PRETRAINED VIDEO MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Three-dimensional medical imaging is typically approached as a collection of specialized problems, with separate architectures, output representations, and training objectives for segmentation, restoration, reconstruction, synthesis, and generation. We ask whether these heterogeneous tasks can instead be addressed through a shared prediction interface built upon a pretrained video model. We introduce **MedGen3D**, a generalist framework that unifies five supervised 3D medical imaging task families within a single latent prediction formulation. To accommodate segmentation, we represent segmentation masks as signed distance fields, converting discrete targets into continuous volumetric representations that can be processed by the same latent prediction pipeline as image-valued targets. A frozen video VAE encodes all inputs and targets, while a single LoRA-adapted video Transformer predicts target latents using the same masked regression objective, without task-specific prediction heads or losses. Across AbdomenAtlas, BraTS 2021, and CT-RATE v2, a single MedGen3D checkpoint performs all five task families within the same framework, reducing restoration and sparse-view reconstruction MAE by 9.7% and 38.6% and improving T1-to-T2 synthesis PSNR by 2.76 dB over the strongest evaluated baselines, while exhibiting task-dependent trade-offs on segmentation and report-conditioned generation. Under matched per-task update budgets, unified training achieves better reported metrics than independently trained single-task counterparts on four of the five task families. These results demonstrate that a pretrained video model can provide a shared representation and prediction interface across substantially different 3D medical imaging tasks, suggesting a path toward generalist volumetric medical models without task-specific prediction architectures.

1 INTRODUCTION

Progress in 3D medical imaging has largely been driven by increasingly specialized models. Strong systems have been developed for anatomical segmentation, image restoration, tomographic reconstruction, cross-modal synthesis, and conditional generation, but these advances have mostly followed separate methodological paths. Models designed for one task are rarely reusable for another: segmentation typically relies on specialized prediction heads and classification- or overlap-based objectives (Isensee et al., 2021); restoration and reconstruction methods are often built around degradation-specific or physics-based formulations (Chen et al., 2017; Jin et al., 2017); cross-modal synthesis is commonly treated as image translation (Zhu et al., 2023); and conditional generation generally introduces its own iterative generative inference procedure (Hamamci et al., 2024b; Guo et al., 2025a). Consequently, addressing a new volumetric imaging problem often requires redesigning not only the model architecture, but also the output representation, optimization objective, and inference pipeline.

This fragmentation raises a more fundamental question: **how much of this task-specific machinery is actually necessary?** Despite their different formulations, these tasks operate on closely related volumetric anatomical signals and can often be viewed as conditional transformations from one 3D representation to another. Recent work has shown that generative models can be adapted for visual prediction (Gabeur et al., 2026; Yang et al., 2026; Wang et al., 2026), while frozen natural-image and video VAEs can provide reusable latent interfaces for medical volumes such as CT (Chen et al., 2026). Together, these developments motivate a more specific question: can heterogeneous 3D medical imaging tasks share one frozen video latent interface and one supervised predictor without

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.

task-specific prediction heads? A central challenge is representation. Continuous image targets, discrete segmentation masks, different observation regimes, and task-dependent conditioning must all be accommodated by the same model while preserving the distinctions required by each task.

We introduce **MedGen3D** to investigate this question. For segmentation, signed distance fields (SDFs) convert discrete masks into continuous volumetric targets that can be processed through the same latent interface as image-valued targets. A frozen WAN2.2 VAE (Team et al., 2025) encodes all inputs and targets into a shared latent space, while a single video DiT, adapted with LoRA (Hu et al., 2022), predicts the corresponding target latent using a common masked regression objective. Text conditioning specifies the desired transformation. For sparse-view reconstruction, the number of projection views is provided as an additional condition; for report-conditioned generation, the position of each 65-slice window within the full volume is encoded when needed. Each 65-slice window is predicted with a single DiT evaluation rather than iterative generative sampling. For tasks requiring full-volume outputs, predictions from overlapping windows are fused to reconstruct the complete volume.

The resulting framework shares the same frozen latent interface, video Transformer, and regression objective across substantially different 3D medical output behaviors. Under paired supervision, direct target-latent regression provides a common prediction formulation for all five task families. For a fixed input, condition, and output grid, MedGen3D produces a deterministic prediction rather than sampling from a full conditional output distribution. This formulation allows conventional prediction tasks and report-conditioned generation to be handled by the same backbone while avoiding iterative sampling at inference time.

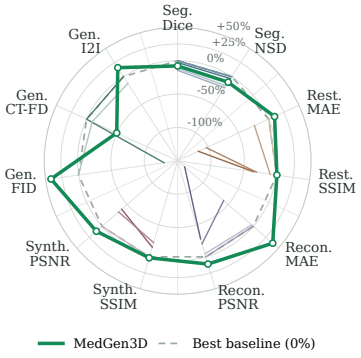

Figure 1: Relative performance across five task families. Values are signed changes from the strongest evaluated baseline (0%); positive is better.

Experiments on AbdomenAtlas, BraTS 2021, and CT-RATE v2 show that a single MedGen3D model can support all five task families despite substantial differences in output type, conditioning, and evaluation protocol. The unified model improves over the strongest evaluated baselines on several restoration, sparse-view reconstruction, and cross-modal synthesis metrics, while segmentation remains below the strongest specialist and report-conditioned generation exhibits metric-dependent trade-offs. We further compare unified training with independently trained single-task models under matched task exposure. When each single-task model receives the same number of task-specific updates as the corresponding task receives during unified training, the unified model achieves better reported metrics on four of the five task families. These results suggest that sharing model parameters across heterogeneous tasks does not necessarily require sacrificing task-level performance, while the remaining gaps in segmentation accuracy and report-conditioned fidelity identify where the unified formulation still falls short of specialized approaches.

Our contributions are threefold:

1. **A unified output representation for heterogeneous 3D medical tasks.** We identify output heterogeneity as a key obstacle to task unification and use signed distance fields to convert discrete segmentation masks into continuous volumetric targets, allowing segmentation and image-valued tasks to share the same latent prediction pipeline.

2. **A shared video-derived prediction interface without task-specific heads.** Building on this common representation, we use a frozen video VAE as a shared input–output latent interface and a single LoRA-adapted video Transformer with a common masked regression objective, enabling five substantially different task families to be handled by the same predictor.

3. **An empirical study of the benefits and limitations of cross-task sharing.** We evaluate the unified model against task-specific methods and matched-exposure single-task controls, showing that parameter sharing can preserve or improve task-level performance across most task families while revealing remaining limitations in segmentation accuracy and report-conditioned generation.

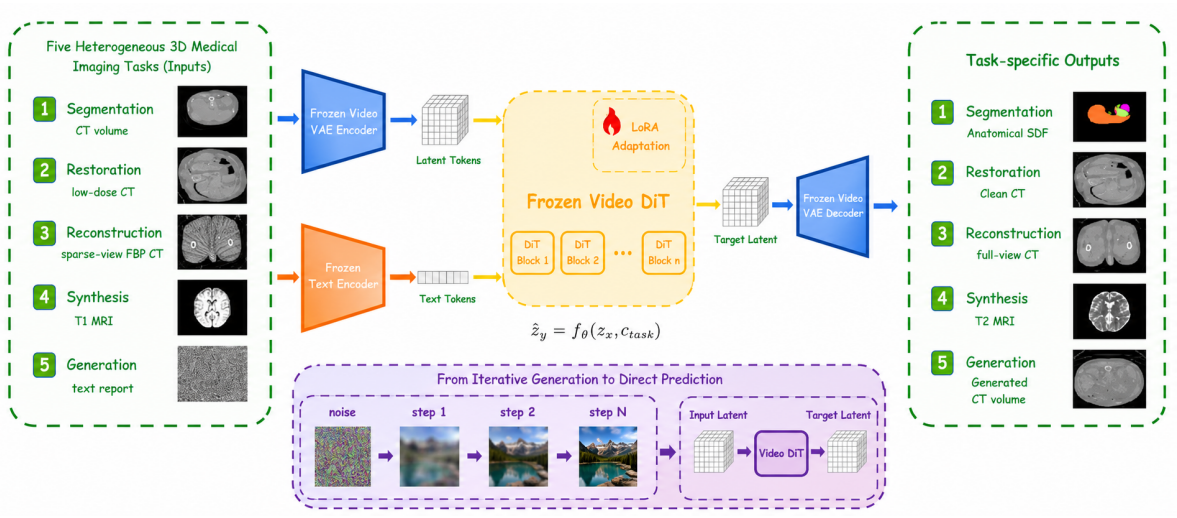

Figure 2: Overview of MedGen3D. Five heterogeneous 3D medical imaging task families are formulated as conditional latent prediction using a shared frozen video VAE interface and a common video DiT backbone. Frozen components are distinguished from the shared trainable LoRA parameters. Sparse-view reconstruction additionally conditions on the projection-view count, while report-conditioned generation encodes the position of each 65-slice window within the full volume. Each window requires a single DiT evaluation rather than iterative sampling. For segmentation, each evaluation predicts the SDF of one prompted organ, which is thresholded to recover a binary mask. For visualization, predictions obtained separately for different organ prompts are combined into the multi-organ example shown in the figure.

2 Related Work

Universal and multi-task medical imaging. Universal segmentation models broaden one predictor across anatomies, datasets, and modalities through partial labels, examples, language, or prompts (Liu et al., 2023; Butoi et al., 2023; Du et al., 2024; Ma et al., 2025; Rokuss et al., 2026), but retain segmentation as the output type. Medical foundation models instead learn transferable or multimodal representations from diverse supervision (Sch"afer et al., 2024; Blankemeier et al., 2024; Wu et al., 2023). MedVerse is closest to our cross-task setting: it performs multiple full-resolution 3D operations through in-context autoregression (Hu et al., 2026). MedGen3D instead casts segmentation, restoration, reconstruction, synthesis, and report-conditioned generation as supervised target-latent prediction with one fixed latent space, predictor, and objective.

Latent representations for medical volumes. Latent modeling reduces the cost of high-resolution volumes. MedVAE trains medical 2D/3D autoencoders that retain clinically relevant information (Varma et al., 2025), whereas Foundation VAE transfers frozen natural-image and video autoencoders to CT (Chen et al., 2026). We study a complementary question: whether one frozen video VAE can encode both inputs and heterogeneous targets for five supervised tasks. Signed distance fields map discrete masks into this continuous interface, allowing a single target-latent predictor to cover both segmentation and image-valued outputs.

Generative models as generalist vision learners. Natural-image generators have recently been adapted into generalist predictors by expressing heterogeneous outputs in a shared visual representation (Gabeur et al., 2026; Yang et al., 2026); GenCeption extends this idea to pretrained video models and feed-forward perception (Wang et al., 2026). Medical diffusion and conditional generation span restoration, inverse problems, MRI synthesis, and text-conditioned CT generation, but generally remain task-specific (Kazerouni et al., 2023; Zhu et al., 2023; Hamamci et al., 2024b; Guo et al., 2025a). MedGen3D brings the generalist formulation to volumetric medical outputs: rather than represent every target as RGB, it uses a frozen video VAE and one video Transformer to predict continuous target latents across distinct acquisition states and conditioning signals.

3 METHOD

MedGen3D addresses three barriers to cross-task sharing: heterogeneous output representations, the compatibility of one frozen latent space with medical signals, and different conditioning requirements. The method uses the same path for every task: encode an input window, condition one shared video DiT, predict the paired target latent, and decode it with the frozen VAE.

3.1 UNIFIED TASK REPRESENTATION

We formulate all five task families through a common conditional volumetric prediction interface,

$$\mathbf{Y} = f(\mathbf{X}, \mathbf{c}), \tag{1}$$

where $\mathbf{X}$ denotes the task-dependent volumetric input, $\mathbf{c}$ specifies the desired transformation together with optional task-specific conditions, and $\mathbf{Y}$ denotes the corresponding target volume. The five tasks are instantiated as

$$f(\mathbf{X}, \mathbf{c}) = \begin{cases} \mathbf{Y}_{\mathrm{SDF}}, & \mathbf{X} = \mathbf{X}_{\mathrm{CT}},\ \mathbf{c} = \mathbf{c}_{\mathrm{seg}}, & \text{Segmentation}, \\ \mathbf{Y}_{\mathrm{CT}}, & \mathbf{X} = \mathbf{X}_{\mathrm{deg}},\ \mathbf{c} = \mathbf{c}_{\mathrm{rest}}, & \text{Restoration}, \\ \mathbf{Y}_{\mathrm{CT}}, & \mathbf{X} = \mathbf{X}_{\mathrm{FBP}},\ \mathbf{c} = (\mathbf{c}_{\mathrm{recon}}, c_{\mathrm{view}}), & \text{Reconstruction}, \\ \mathbf{Y}_{\mathrm{T2}}, & \mathbf{X} = \mathbf{X}_{\mathrm{T1}},\ \mathbf{c} = \mathbf{c}_{\mathrm{syn}}, & \text{Synthesis}, \\ \mathbf{Y}_{\mathrm{CT}}, & \mathbf{X} = \mathbf{X}_{\mathrm{gen}},\ \mathbf{c} = \mathbf{c}_{\mathrm{report}}, & \text{Generation}. \end{cases} \tag{2}$$

Segmentation is represented as continuous signed distance field (SDF) prediction rather than discrete mask classification. Given an input CT volume and an organ condition, the model predicts the corresponding organ SDF, from which the binary segmentation is recovered by thresholding at zero. Restoration maps a degraded CT volume to its clean counterpart. Sparse-view reconstruction takes an FBP-reconstructed CT as input and additionally receives the acquisition setting c_{view}. Cross-modal synthesis maps spatially aligned T1-weighted MRI to T2-weighted MRI. For report-conditioned generation, where no spatially corresponding source image exists, $\mathbf{X}_{\mathrm{gen}}$ is a deterministic, sample-specific volumetric carrier independent of image intensities and report content. Its grid fixes the requested output extent, while the radiology report provides the paired text condition for the target CT.

Under this formulation, task-specific differences are absorbed into the construction of $\mathbf{X}$, $\mathbf{c}$, and $\mathbf{Y}$, while the underlying computational abstraction remains unchanged. In particular, all five task families produce a volumetric target through the same prediction interface, eliminating the need for task-specific output heads. Details of task-specific input construction and preprocessing are provided in Appendix B.

3.2 FROZEN VIDEO VAE AS A SHARED LATENT INTERFACE

Latent compression makes full-resolution volumetric prediction tractable, but an encoder that erases small structures or cannot represent SDFs would undermine the unified interface. We therefore use the pretrained WAN2.2 video VAE for every input and target, keep it completely frozen, and test its round-trip fidelity against a medical-domain VAE in Sec. 4.4. A video VAE is structurally relevant because both videos and medical volumes organize 2D spatial content along an ordered third axis. We exploit this shared structure without interpreting anatomical depth as time.

Let E_{VAE} and D_{VAE} denote the pretrained video encoder and decoder. For a task input $\mathbf{X}$ and target $\mathbf{Y}$, we obtain

$$\mathbf{z}_X = E_{\mathrm{VAE}}(\mathbf{X}), \qquad \mathbf{z}_Y = E_{\mathrm{VAE}}(\mathbf{Y}), \tag{3}$$

and the model predicts a target latent $\hat{\mathbf{z}}_Y$. The final volumetric output is recovered using the same frozen decoder,

$$\hat{\mathbf{Y}} = D_{\mathrm{VAE}}(\hat{\mathbf{z}}_Y). \tag{4}$$

The same encoder–decoder pair is shared across all five task families. It therefore serves as a common interface for heterogeneous imaging modalities such as CT and MRI, different input states such as clean, degraded, and sparse-view reconstructed volumes, and structurally different prediction targets

such as image intensities and continuous SDFs. No medical-specific modification or fine-tuning is applied to the pretrained VAE itself.

This design separates representation from task specialization: heterogeneous medical signals are first expressed in a common pretrained latent space, while the downstream DiT learns the conditional transformation between input and target latents. We empirically evaluate the compatibility of this frozen latent interface across modalities, input states, and target types in Sec. 4.4.

3.3 Shared Video DiT with LoRA Adaptation

To keep adaptation parameter-efficient while retaining one shared backbone, MedGen3D freezes the original WAN2.2 video DiT and inserts LoRA modules (Hu et al., 2022) into its linear layers. The DiT and LoRA parameters are shared across all tasks. Apart from the lightweight view-count and generation-window conditioners described below, MedGen3D has no task-specific backbone or prediction head.

The encoded input latent $\mathbf{z}_X$ is tokenized by the pretrained DiT patch embedding and processed directly by the shared backbone. Text specifies the desired task transformation, with additional conditioning introduced only when required by the task semantics. Sparse-view reconstruction receives the normalized acquisition setting c_{view}, allowing the same reconstruction model to condition on the measurement configuration. Report-conditioned generation uses the frozen pretrained text encoder to convert the radiology report into a text-conditioning embedding. A lightweight position condition additionally encodes the normalized axial start and extent of each generation window. The view ratio and generation-window position are represented using sinusoidal Fourier features followed by two-layer MLPs, whose outputs are added to the DiT conditioning embedding. These conditions modify the behavior of the common backbone without changing its output representation or introducing task-specific prediction heads.

This backbone- and LoRA-sharing strategy is central to the five-task setting considered here: the tasks differ in their inputs, targets, and conditioning, but the actual latent transformation is performed by one shared pretrained Transformer with one common set of LoRA adaptation parameters. Sensitivity to block-wise LoRA removal is analyzed separately in Appendix G.

3.4 Unified Feed-Forward Training

WAN2.2 synthesizes video through an iterative Flow Matching (Lipman et al., 2023) trajectory. Paired supervision instead provides a direct target in latent space, including for report–CT pairs. We use this supervision to train a deterministic predictor rather than retain the pretrained model's iterative sampling procedure. This modeling choice supplies one prediction rule across tasks; it is not intended to model the full conditional distribution of plausible outputs.

We instead repurpose the pretrained video DiT as a direct latent predictor. The original timestep-conditioning interface is retained for architectural compatibility but fixed to $t = 0$. Given an input latent $\mathbf{z}_X$ and condition $\mathbf{c}$, the model directly predicts the corresponding target latent with one Transformer evaluation per 65-slice window,

$$\hat{\mathbf{z}}_Y = F_\theta(\mathbf{z}_X, t = 0, \mathbf{c}) \approx \mathbf{z}_Y, \tag{5}$$

where θ denotes the trainable LoRA, view-conditioning, and position-conditioning parameters. The pretrained VAE, text encoder, and original DiT weights remain frozen.

Because all task targets are represented in the same latent space, the same regression objective can be applied to every task family. We optimize a masked mean-squared error between the predicted and target latents,

$$\mathcal{L}_{\text{reg}} = \text{MaskedMSE}(\hat{\mathbf{z}}_Y, \mathbf{z}_Y; \mathbf{M}_z)\,, \tag{6}$$

where $\mathbf{M}_z$ excludes padded latent positions from optimization. Unlike conventional multi-task systems that combine heterogeneous task-specific objectives, the same loss is used unchanged for segmentation, restoration, reconstruction, synthesis, and generation.

During multi-task training, all five task families remain active and share the same architecture and regression objective; the task-sampling schedule is described in Section 4.1.

Table 1: **Comparison with task-specific methods across five task families.** On matched test cohorts, MedGen3D leads all reported restoration, reconstruction, and synthesis metrics, while trailing the strongest segmentation specialist and showing mixed generation results. A shared latent predictor therefore supports strong image-to-image performance, with task- and metric-dependent limits to replacing specialized systems.

Task family	Method	Task-specific metrics			
		Dice ↑	NSD@3 mm ↑	ASSD† ↓	Empty ↓
Anatomical segmentation	MedNeXt (LCC) (Roy et al., 2023)	0.892	0.911	2.54	**0.0**
	TotalSegmentator (Wasserthal et al., 2022)	0.917	0.943	1.97	1.0
	nnU-Net v2 (Isensee et al., 2021)	**0.918**	**0.945**	**1.63**	**0.0**
	3D U-Net (Çiçek et al., 2016)	0.786	0.758	8.21	2.0
	Residual 3D U-Net	0.812	0.792	6.47	1.0
	MedGen3D	0.846	0.851	3.34	1.5
		MAE_{HU} ↓	$RMSE_{HU}$ ↓	PSNR ↑	SSIM(2D) ↑
CT restoration	CTformer (Wang et al., 2023)	58.23	80.53	28.13	0.569
	CoreDiff (Gao et al., 2024)	54.70	74.44	29.03	0.606
	RED-CNN (Chen et al., 2017)	33.30	48.22	32.57	0.787
	Residual 2D U-Net	26.98	42.94	33.47	0.875
	MedGen3D	**24.37**	**40.46**	**34.04**	**0.877**
		MAE_{HU} ↓	$RMSE_{HU}$ ↓	PSNR ↑	SSIM(2D) ↑
Sparse-view CT reconstruction	FBP	129.99	202.56	19.96	0.341
	ASD-POCS (Qiao, 2021)	87.48	157.55	22.12	0.566
	FBPConvNet (Jin et al., 2017)	55.07	83.26	27.77	0.725
	Residual 2D U-Net	56.80	97.21	26.43	0.703
	MedGen3D	**33.83**	**58.51**	**30.90**	**0.802**
		MAE ↓	PSNR ↑	SSIM(2D) ↑	–
T1-to-T2 MRI synthesis	YODA (Rassmann et al., 2026)	0.121	15.08	0.783	–
	SynDiff (Ozbey et al., 2023)	0.106	16.61	0.717	–
	Residual 2D U-Net	0.018	24.76	0.922	–
	MedGen3D	**0.013**	**27.52**	**0.936**	–
		FID ↓	CT-CLIP Fréchet ↓	I2I ↑	–
Report-conditioned CT generation	GenerateCT (Hamamci et al., 2024b)	32.002	**0.413**	45.33	–
	Text2CT (Guo et al., 2025a)	13.950	0.451	39.81	–
	MedGen3D	**8.130**	0.618	**52.51**	–

4 Experiments

We evaluate MedGen3D on five heterogeneous tasks: anatomical segmentation, image restoration, sparse-view CT reconstruction, cross-modal MRI synthesis, and report-conditioned CT generation. We compare the unified model with task-specific methods, test the frozen latent interface against a medical VAE, and measure where the shared DiT depends on medical adaptation. Zero-shot external-dataset results are provided in Appendix E.

4.1 Experimental Setup

We evaluate MedGen3D on five 3D medical imaging tasks using fixed patient-disjoint training, validation, and held-out test sets. Dataset details and task-specific input–target construction are given in Appendix B.

The model is initialized from WAN2.2 (Team et al., 2025) with the video VAE, text encoder, and original Transformer frozen; only LoRA and lightweight view- and position-conditioning modules are optimized. We train with a global batch size of 4. We sample the five tasks uniformly for the first 60k steps. Because segmentation remained substantially less optimized than the other task families under uniform sampling, we increased its sampling frequency to 3:1:1:1:1 from 60k–80k steps and 6:1:1:1:1 from 80k–86k steps, while keeping all five tasks active throughout. The schedule modifies task exposure only; the shared architecture and training objective remain unchanged. Full implementation details are provided in Appendix C.

Table 2: **Unified versus single-task training at matched task exposure.** The unified model improves both reported metrics for segmentation, reconstruction, synthesis, and generation; single-task training performs better on restoration. Joint training can therefore benefit several heterogeneous tasks without increasing their individual update budgets, although the benefit is not uniform across tasks.

Task	Metrics	Unified	Single-task	Updates
Segmentation	Dice / NSD	.846 / .851	.823 / .815	24.17k
Restoration	MAE / PSNR	24.37 / 34.04	23.43 / 34.35	15.46k
Reconstruction	MAE / PSNR	33.83 / 30.90	34.39 / 30.76	15.46k
Synthesis	MAE / PSNR	.013 / 27.52	.015 / 27.05	15.46k
Generation	FID / I2I	8.13 / 52.51	9.40 / 51.20	15.46k

Although all five tasks share the same latent regression objective, evaluation follows the standard metrics of each task family. We report Dice, NSD, ASSD, and empty prediction rate for segmentation; MAE, RMSE, PSNR, and SSIM for restoration and reconstruction; and MAE, PSNR, and SSIM for synthesis. Generation is evaluated using MAISI (Guo et al., 2025b) 2.5D RadImageNet (Mei et al., 2022) FID (Heusel et al., 2017), CT-CLIP (Hamamci et al., 2024a) Fréchet distance, and CT-CLIP reference-to-generated-volume similarity (I2I) in Table 1. All metrics are computed on the held-out test sets at the case level unless otherwise specified.

4.2 Comparison with Task-Specific Models

A common prediction interface is useful only if it retains the capabilities of task-specific systems. We therefore compare one MedGen3D checkpoint against dedicated methods for all five task families, using the same fixed test cases, preprocessing, and metric implementations within each task. Table 1 tests where a shared latent representation and regression objective remain competitive, and where specialization still provides an advantage.

MedGen3D achieves the best reported values on every evaluated metric for restoration, sparse-view reconstruction, and T1-to-T2 synthesis. Relative to the strongest evaluated baseline, restoration MAE decreases by 9.7% (26.98 to 24.37 HU, versus residual 2D U-Net), and reconstruction MAE by 38.6% (55.07 to 33.83 HU, versus FBPConvNet). Synthesis PSNR increases from 24.76 to 27.52 dB over residual 2D U-Net, alongside lower MAE and higher SSIM. The agreement across intensity-error and structural-similarity metrics supports the shared formulation for these image-to-image tasks.

Segmentation exposes a remaining gap: MedGen3D reaches .846 Dice and .851 NSD, compared with .918 and .945 for nnU-Net v2, although its Dice exceeds both evaluated 3D U-Net variants (.786 and .812). Converting masks to SDFs makes segmentation compatible with the common interface, but does not by itself ensure equally accurate binary masks or boundaries. Latent compression and the mismatch between SDF regression and mask-based evaluation are possible contributors; this comparison does not isolate their effects. The matched single-task comparison below examines whether joint training contributes to the gap.

Report-conditioned generation shows a different trade-off. MedGen3D improves FID to 8.130 from Text2CT's 13.950 and I2I to 52.51 from GenerateCT's 45.33, but its CT-CLIP Fréchet distance is higher (.618 versus .413 for GenerateCT). Thus, improved RadImageNet feature-distribution matching and paired-volume similarity coexist with weaker distribution matching in CT-CLIP space. Together, the results establish strong image-to-image performance under one shared predictor, while identifying segmentation accuracy and metric-dependent generation quality as the remaining limitations.

4.3 Unified versus Single-Task Training

Comparisons with task-specific systems do not separate the effect of joint training from differences in architecture and initialization. We therefore compare MedGen3D with independently trained single-task counterparts, keeping the architecture, initialization, LoRA configuration, data splits, and evaluation protocol unchanged. Each single-task model receives approximately the same number of task-specific updates as its task receives in the unified run: 24.17k for segmentation and 15.46k for each other task. This controls exposure to the evaluated task, while the unified model additionally

Table 3: **Frozen-VAE compatibility across signal types.** On 20 matched cases without adaptation, the video VAE exceeds 3D MedVAE in CT/MRI fidelity, sparse-view residual preservation (higher correlation and lower error), and liver-SDF Dice/NSD. These results support its use as a shared latent interface for both image-valued and geometric targets.

Category	Signal	Metrics	Video VAE	MedVAE
Modality	CT	PSNR / SSIM	**36.47 / .929**	34.02 / .886
Modality	T1 MRI	PSNR / SSIM	**29.69 / .964**	27.13 / .937
Modality	T2 MRI	PSNR / SSIM	**29.97 / .958**	27.52 / .929
Input state	Sparse-view FBP CT	Residual corr. / error	**.762 / .670**	.660 / .794
Target type	Liver SDF	Dice / NSD	**.994 / .869**	.991 / .815

Table 4: **Zero-shot transfer to external cohorts.** Without external adaptation, the same MedGen3D checkpoint surpasses the strongest evaluated specialist on the primary metric for restoration, reconstruction, synthesis, and generation across fixed 200-case cohorts, while trailing on segmentation. The shared predictor therefore transfers competitively across image-valued tasks, with segmentation remaining a limitation.

Task / cohort	Metric	Strongest specialist	MedGen3D	Better
Segmentation / DAP-Atlas	Dice ↑	nnU-Net v2: .917	.801	Specialist
Restoration / DAP-Atlas	MAE_{HU} ↓	RED-CNN: 27.23	**19.93**	MedGen3D
Reconstruction / DAP-Atlas	MAE_{HU} ↓	FBPConvNet: 56.45	**29.37**	MedGen3D
Synthesis / BraTS-PEDs	MAE ↓	YODA: .195	**.108**	MedGen3D
Generation / BIMCV-R	FID ↓	Text2CT: 41.41	**8.38**	MedGen3D

learns from the other four tasks; it does not match the total updates of the unified model to those of each individual single-task model.

Table 2 shows better values on both reported metrics for four of the five tasks. Segmentation Dice/NSD increase from .823/.815 to .846/.851, indicating that joint training improves segmentation within this formulation despite its remaining gap to specialist systems. Reconstruction MAE decreases from 34.39 to 33.83 HU, synthesis PSNR increases from 27.05 to 27.52 dB, and generation FID decreases from 9.40 to 8.13 with higher I2I. Restoration is the exception: independent training achieves 23.43 versus 24.37 HU MAE and 34.35 versus 34.04 dB PSNR.

These gains support joint training at fixed per-task exposure, but the restoration result shows that outperforming task-specific baselines does not guarantee a benefit from sharing. Conversely, segmentation improves through joint training while remaining below specialists. The two comparisons therefore distinguish system-level competitiveness from the task-dependent effect of joint training.

4.4 Frozen Video VAE as a Shared Latent Interface

A shared frozen encoder–decoder must preserve the distinct signals required by all five tasks: medical image intensities, degraded observations, and SDF geometry. We therefore test whether the video VAE can accommodate these signals without medical-domain adaptation. We compare its encode–decode round trip with 3D MedVAE on the same 20 cases under matched preprocessing, keeping both models frozen. This evaluates the latent interface separately from the learned DiT predictor.

Table 3 shows higher reconstruction fidelity for CT and both MRI contrasts, including 36.47 versus 34.02 dB CT PSNR. For sparse-view CT, higher residual correlation (.762 versus .660) and lower residual error (.670 versus .794) indicate better preservation of the structured corruption present in the input; these metrics do not measure recovery of clean CT. For liver SDFs, .994 Dice and .869 NSD show that the interface also retains geometric targets, with higher boundary fidelity than MedVAE (.815 NSD). Together, these results support one frozen interface for the tested signal types, while downstream prediction accuracy remains a separate question.

4.5 Zero-Shot Transfer to External Cohorts

Performance on held-out cases from the training datasets does not establish transfer to other cohorts. We therefore evaluate the same MedGen3D checkpoint on fixed 200-case external cohorts: DAP-Atlas for the three CT tasks, BraTS-PEDs for MRI synthesis, and BIMCV-R for report-conditioned

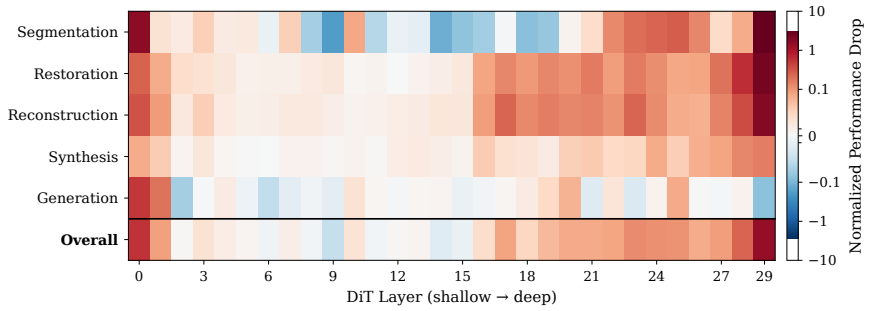

Figure 3: **Layer-wise LoRA sensitivity.** Disabling one block's LoRA update at a time on 40 held-out samples per task gives the largest overall drops at blocks 0 and 29, while generation is most sensitive to blocks 0–1. Positive values denote degradation and negative values improvement. These differing profiles indicate task-dependent reliance on the shared LoRA adaptation.

generation. MedGen3D and all evaluated specialists are selected using only in-distribution validation data and receive no external adaptation. Within each task, methods share the cases, preprocessing, and metrics. The CT corruption and 18-view reconstruction settings are unchanged, so this tests cohort transfer under fixed degradation protocols.

Table 4 shows that MedGen3D outperforms the strongest evaluated specialist on the primary metric for each of the four image-valued tasks. Restoration and reconstruction MAE decrease from 27.23 to 19.93 HU and from 56.45 to 29.37 HU, respectively; synthesis MAE decreases from .195 to .108, and generation FID from 41.41 to 8.38. Segmentation remains weaker, with .801 Dice versus .917 for nnU-Net v2. Thus, a shared predictor can remain competitive beyond its training cohorts, but its transfer performance is task-dependent. This comparison does not isolate the contributions of pretraining, model scale, or upstream data overlap. Full metrics and cohort details are in Appendix E.

4.6 Sensitivity to LoRA Adaptation

Sharing a backbone does not imply that every task relies equally on its adapted layers. To locate this dependence, we disable one DiT block's LoRA update at a time while keeping all other parameters fixed, and evaluate the same 40 held-out samples per task. We average direction-aware relative performance drops across each task's metrics; positive values indicate degradation, and Overall averages the five task scores.

Figure 3 reveals distinct depth profiles. Blocks 0 and 29 produce the largest overall drops. Segmentation is highly sensitive to both, while restoration and reconstruction also depend on a broader range of deeper blocks. Synthesis shows smaller effects that increase toward the final blocks. Generation is most sensitive to blocks 0–1 and shows mixed, weaker effects at greater depth, including an improvement when block 29 is disabled. Thus, the tasks rely on different parts of the shared adaptation, and the overall profile should not be read as a universal layer-importance ranking. This probe characterizes the trained model; identifying an optimal LoRA placement would require retraining. The full numerical matrix is in Appendix G.

5 Conclusion

We presented **MedGen3D**, a unified latent framework for five 3D medical imaging task families. Signed distance fields align masks with the common video VAE, LoRA-adapted Transformer, and regression objective. Across datasets, MedGen3D is strong on continuous volumetric prediction and outperforms single-task counterparts on four of five tasks under matched exposure. This supports pretrained video representations as a shared prediction interface; remaining segmentation and report-conditioned generation gaps motivate boundary-aware modeling and stronger semantic conditioning.

AI USE STATEMENT

We used generative AI tools, including ChatGPT and Codex, to assist with several aspects of the research process. ChatGPT was used primarily for literature understanding, discussion and analysis of research methodology and experimental design, interpretation of experimental results, and improvement of the clarity and readability of the manuscript. Codex was used primarily to assist with scientific software development, including code implementation, modification, debugging, and experimental pipeline development.

All AI-assisted outputs were reviewed by the authors. AI-generated or modified code was checked and validated through inspection and actual execution. Experimental results and scientific conclusions were based on the authors' own experiments and analysis rather than AI-generated predictions. The authors take full responsibility for the final content of this work, including all claims, experimental results, and materials produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses existing medical imaging datasets and does not involve new data collection or direct interaction with human participants. The datasets are used solely for research purposes in accordance with their respective access conditions and usage requirements. MedGen3D is evaluated as a research model and is not intended for direct clinical deployment or diagnostic decision-making. Generated or predicted medical volumes should not be interpreted as clinically validated patient-specific findings. We report both strengths and limitations across task families to avoid overstating the reliability of the unified model in settings where specialized clinical systems may remain preferable.

REPRODUCIBILITY STATEMENT

We provide the data splits, preprocessing procedures, task-specific input–target construction, evaluation protocols, model configuration, optimization hyperparameters, multi-task sampling schedule, and inference settings required to reproduce the reported experiments. All dataset splits are patient-disjoint and fixed before final evaluation. Preprocessing and simulated corruptions are generated deterministically and reused across methods where applicable. The appendix further specifies window sampling, random seeds, metric implementations, VAE analyses, single-task controls, external-dataset evaluation, and LoRA sensitivity experiments. These details are intended to make the training and evaluation pipeline reproducible and to enable direct comparison with the reported results.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, et al. Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512*, 2024.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023.

Hu Chen, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network (red-cnn). *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, 2017. doi: 10.1109/TMI.2017.2715284.

Qi Chen, Shuhan Ding, Yu Gu, Nan Liu, Jiang Bian, Alan L. Yuille, Zongwei Zhou, and Jingjing Fu. Foundation vae for ct reconstruction, augmentation, and generation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. BIMCV-R: A landmark dataset for 3d CT text-image retrieval. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 124–134, 2024. doi: 10.1007/978-3-031-72120-5_12.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432, 2016. doi: 10.1007/978-3-319-46723-8_49.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. In *Advances in Neural Information Processing Systems*, 2024.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026. doi: 10.48550/arXiv.2604.20329.

Qi Gao, Zilong Li, Junping Zhang, Yi Zhang, and Hongming Shan. Corediff: Contextual error-modulated generalized diffusion model for low-dose ct denoising and generalization. *IEEE Transactions on Medical Imaging*, 43(2):745–759, 2024. doi: 10.1109/TMI.2023.3320812.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

Pengfei Guo, Can Zhao, Dong Yang, Yufan He, Vishwesh Nath, Ziyue Xu, Pedro R. A. S. Bassi, Zongwei Zhou, Benjamin Simon, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. *arXiv preprint arXiv:2505.04522*, 2025a.

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 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4430–4441, 2025b. doi: 10.1109/WACV61041.2025.00435.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024a.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, Chinmay Prabhakar, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision – ECCV 2024*, 2024b.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jiesi Hu, Jianfeng Cao, Yanwu Yang, Chenfei Ye, Yixuan Zhang, Hanyang Peng, and Ting Ma. Medverse: A universal model for full-resolution 3d medical image segmentation, transformation and enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9): 4509–4522, 2017. doi: 10.1109/TIP.2017.2713099.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, et al. The brain tumor segmentation (BraTS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). *arXiv preprint arXiv:2305.17033*, 2023.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. doi: 10.1016/j.media.2023.102846.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21152–21164, 2023.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315, 2022.

Muzaffer Ozbey, Onat Dalmaz, Salman UH Dar, Hasan A. Bedel, Saban Ozturk, Alper Gungor, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023. doi: 10.1109/TMI.2023.3290149.

Zhiwei Qiao. A simple and fast asd-pocs algorithm for image reconstruction. *Journal of X-Ray Science and Technology*, 29(3):491–506, 2021. doi: 10.3233/XST-210858.

Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *arXiv preprint arXiv:2305.09666*, 2023.

Sebastian Rassmann, David Kugler, Christian Ewert, and Martin Reuter. Regression is all you need for medical image translation. *IEEE Transactions on Medical Imaging*, 2026. doi: 10.1109/TMI.2025.3650412.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, and Klaus Maier-Hein. Voxtell: Free-text promptable universal 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37538–37557, 2026.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023.

Raphael Sch"afer, Till Nicke, Henning H"ofener, Annkristin Lange, Dorit Merhof, Friedrich Feuerhake, Volkmar Schulz, Johannes Lotz, and Fabian Kiessling. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 4(7):495–509, 2024. doi: 10.1038/s43588-024-00662-z.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Maya Varma, Ashwin Kumar, Rogier van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay Chaudhari. Medvae: Efficient automated interpretation of medical images with large-scale generalizable autoencoders. *arXiv preprint arXiv:2502.14753*, 2025.

Dayang Wang, Fenglei Fan, Zhan Wu, Rui Liu, Fei Wang, and Hengyong Yu. Ctformer: Convolution-free token2token dilated vision transformer for low-dose ct denoising. *Physics in Medicine and Biology*, 68(6):065012, 2023. doi: 10.1088/1361-6560/acc000.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Video generation models are general-purpose vision learners. *arXiv preprint arXiv:2607.09024*, 2026. doi: 10.48550/arXiv.2607.09024. ECCV 2026.

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.

Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d and 3d medical data. *arXiv preprint arXiv:2308.02463*, 2023.

Timing Yang, Jinrui Yang, Xinlong Li, Yuhan Wang, Haoran Li, Yanqing Liu, Guoyizhe Wei, Jixuan Ying, Chen Wei, Rama Chellappa, Yuyin Zhou, Cihang Xie, Alan Yuille, and Feng Wang. Let rgb be the language of vision. *arXiv preprint arXiv:2607.12450*, 2026. doi: 10.48550/arXiv.2607.12450.

Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 592–601, 2023. doi: 10.1007/978-3-031-43999-5_56.