SciFusion: Compositional Representations for Physics-Guided 3D Scientific Image Generation

Luoxi Wang[1,2] **Alessio Quercia**[2] **Hanno Scharr**[2,†] **Arya Bangun**[2,†]

[1]Princeton University, Princeton, NJ, USA
[2]IAS-8: Data Analytics and Machine Learning,
Forschungszentrum Jülich, Jülich, Germany

luoxi.wang@princeton.edu
{a.quercia,h.scharr,a.bangun}@fz-juelich.de

Abstract

Existing 3D generative models for scientific imaging are typically trained separately for each modality, limiting knowledge transfer across heterogeneous imaging domains and requiring substantial modality-specific data. We present SciFusion, a physics-guided compositional latent diffusion framework that learns a unified representation for 3D scientific image generation across modalities. SciFusion factorizes each modality into reusable domain, object, and contrast representations, allowing shared structure to transfer across MRI, CT, and electron microscopy (EM) while preserving modality-specific characteristics, and enabling zero-shot generation of unseen modality compositions without target-modality training. To respect modality-dependent image formation, we incorporate domain-specific physics guidance based on MRI power-spectrum statistics, CT Radon projections, and EM diffraction patterns. Experiments across 15 medical and scientific imaging modalities show that SciFusion improves generation quality under unified training and successfully synthesizes unseen modality compositions. These results suggest that compositional, physics-grounded representations provide a scalable framework for transferable 3D scientific image generation.

1 Introduction

Three-dimensional scientific imaging is central to biomedical research and scientific discovery, spanning modalities such as magnetic resonance imaging (MRI), computed tomography (CT), and electron microscopy (EM). Large-scale 3D datasets remain difficult to collect due to costly acquisition, specialized equipment, and privacy constraints. Recent diffusion models enable high-quality volumetric generation [9, 23, 20, 3], but existing methods are typically specialized to individual datasets or acquisition settings, limiting knowledge transfer across modalities. Scientific imaging is naturally compositional: modalities differ along factors such as imaging domain, anatomical object, and imaging contrast. Treating each modality as an independent category limits generalization to unseen combinations, while naively merging heterogeneous modalities can obscure modality-specific characteristics. A unified model should instead capture shared structure while explicitly representing these factors.

We present SciFusion, a physics-guided compositional latent diffusion framework for unified 3D scientific image generation across MRI, CT, and EM. SciFusion factorizes modality information

[†]These authors contributed equally as senior authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences Workshop.

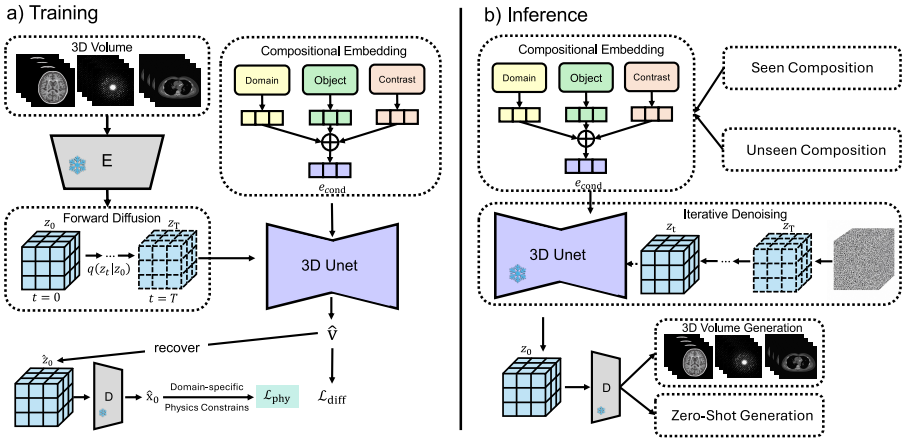

Figure 1: Overview of SciFusion. (a) Two-stage training with shared latent learning, compositional conditioning, and physics guidance. (b) Compositional inference for seen and zero-shot modalities.

into reusable domain, object, and contrast embeddings while incorporating domain-specific physics guidance. Our contributions are:

- A compositional conditioning strategy for generation across heterogeneous imaging data.
- Zero-shot generation of unseen modality combinations by recombining domain, object, and contrast factors without target-modality training data.
- Domain-specific physics guidance tailored to MRI, CT, and EM.
- Evaluation across 15 imaging modalities, demonstrating improved generation quality and compositional generalization.

2 Method

SciFusion builds on latent diffusion [20] and MedLoRD [21] for efficient high-resolution 3D image synthesis (Fig. 1). Training has two stages: a VQGAN is first pretrained on the integrated multi-modal dataset without modality conditioning to learn a shared latent space across heterogeneous imaging domains (Sec. 2.1); a conditional 3D latent diffusion model is then trained on frozen VQGAN latents using compositional conditioning (Sec. 2.2) and domain-specific physics guidance (Sec. 2.3). Training details are in the supplementary material. At inference, recombining learned conditions enables both seen and zero-shot generation of unseen modality combinations.

2.1 Latent Diffusion Framework

Each volume $\mathbf{x} \in \mathbb{R}^{C\times S\times N\times N}$ is encoded by a frozen, pretrained VQGAN as $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{C\times S\times L\times L}$ ($L \ll N$, depth dimension preserved), with reconstruction $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_0)$. A conditional 3D U-Net f_θ, shared across all modalities, predicts the v-prediction target [20] from noisy latent $\mathbf{z}_t$, timestep t, and compositional embedding $\mathbf{e}_{\text{cond}}$ (Sec. 2.2) via $\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0,t,\boldsymbol{\epsilon}}\|\mathbf{v}_t - \hat{\mathbf{v}}_\theta\|_2^2$; generation across domains is controlled entirely by the conditioning mechanism described next.

2.2 Compositional Conditioning Mechanism

Conventional multimodal diffusion models typically assign an independent learnable embedding to each modality [26]; while this can distinguish modalities seen during training, it treats each as an unrelated category without exploiting shared structure across heterogeneous scientific images.

We instead represent each modality as a composition of three semantic factors—imaging domain, anatomical/structural object, and imaging contrast—i.e., $\mathbf{c} = (c_{\text{dom}}, c_{\text{obj}}, c_{\text{con}})$, via three independent learnable embedding tables $\mathcal{E}_{\text{dom}}$, $\mathcal{E}_{\text{obj}}$, $\mathcal{E}_{\text{con}}$ yielding $\mathbf{e}_k = \mathcal{E}_k(c_k) \in \mathbb{R}^{d_{\text{cond}}}$, where d_{cond} is the conditioning dimension. These are combined via element-wise addition into a single conditioning token $\mathbf{e}_{\text{cond}} = \mathbf{e}_{\text{dom}} + \mathbf{e}_{\text{obj}} + \mathbf{e}_{\text{con}} \in \mathbb{R}^{1\times d_{\text{cond}}}$, injected into the U-Net's cross-attention layers. This factorization captures both modality-specific and shared structure and enables compositional zero-

shot generation: unseen modality combinations can be constructed directly from factors observed during training, without any additional modality-specific embedding.

2.3 Physics-Informed Training Objective

Given the predicted noise, we first recover the clean latent estimate $\hat{\mathbf{z}}_0$ and decode it into a predicted volume $\hat{\mathbf{x}}_0$. For each sample, a domain-dependent differentiable operator $\mathcal{P}_d$ is applied to both the reference and predicted volumes:

$$\mathcal{L}_{\text{phys}} = \frac{1}{B} \sum_{b=1}^{B} \left\| \mathcal{P}_{d_b}(\mathbf{x}_b) - \mathcal{P}_{d_b}(\hat{\mathbf{x}}_{0,b}) \right\|_1 , \tag{1}$$

where d_b denotes the imaging domain of the b-th sample, with $\mathcal{P}_d$ instantiated as one of $\mathcal{P}_{\text{MRI}}$, $\mathcal{P}_{\text{CT}}$, or $\mathcal{P}_{\text{EM}}$ below. This formulation allows physical constraints to be applied within the same mixed-domain training batch.

Since MRI is acquired in Fourier space, we constrain the slice-wise spatial-frequency distribution via $\mathcal{P}_{\text{MRI}}(\mathbf{x}) = \log_{10}(1 + |\mathcal{F}_{2\text{D}}(\mathbf{x})|^2)$. For CT, we enforce sinogram-domain consistency using a differentiable Radon transform [8, 28], $\mathcal{P}_{\text{CT}}(\mathbf{x}) = \log_{10}(1 + |\mathcal{R}(\mathbf{x})|^2)$, where $\mathcal{R}$ denotes the Radon transform. For EM, electron wave propagation through the volume is modeled via the multislice method with plane-wave illumination [5, 15] (full slice-wise recursion in the supplementary material), and the diffraction representation is computed from the far-field intensity after propagation through all S slices: $\mathcal{P}_{\text{EM}}(\mathbf{x}) = \log_{10}(1 + |\mathcal{F}_{2D}(\psi_S)|^2)$. The overall training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}$, applied only during training and not altering the inference procedure; $\lambda_{\text{phys}} = 5 \times 10^{-4}$ is used for all experiments, selected via the sweep detailed in the supplementary material (Sec. D).

3 Results

Table 1: Ablation study of SciFusion across four configurations: single-modality models, flattened multimodal conditioning, compositional conditioning, and the full model with physics guidance.

Method	# Models	Comp.	Physics	I3D-FVD ↓	VMAE-FVD ↓	JEDI ↓
Single Modality	15	–	–	13.07	35.93	2.62
Flat Multimodal	1	–	–	11.80	28.63	2.36
Comp. Multimodal	1	✓	–	11.10	27.30	2.35
Ours	1	✓	✓	**9.90**	**23.51**	**1.80**

3.1 Dataset and Evaluation

We construct a 15-modality volumetric benchmark spanning MRI, CT, and EM to evaluate unified multimodal generation and compositional generalization. MRI (knee, brain, prostate, breast) is from NYU fastMRI [16, 29]; CT (head, chest, abdomen) is from RSNA [6] and the Medical Segmentation Decathlon [22]; EM (crystals, apoferritin, hemoglobin) is from the Materials Project [13] and PDB [4]. Training samples are unevenly distributed across modalities, with no explicit rebalancing during unified training, resulting in more frequent sampling of data-rich modalities. For zero-shot evaluation, knee T1 (MRNet [24, 1]) and brain PD (IXI [12]) are held out, with their constituent factors observed during training. Generation quality is evaluated using FVD, KID, and JEDI [18, 2, 25] on both I3D and VideoMAE backbones. Following prior work on volumetric medical image generation [7, 27], we evaluate generation quality using FVD, KID, and JEDI [18, 2, 25] with I3D and VideoMAE features. However, these video-based metrics may not fully capture 3D spatial consistency and anatomical or structural fidelity. See Sec. A of the supplementary material for full dataset construction.

3.2 Effectiveness of Compositional Conditioning and Physics Guidance

We first ablate the conditioning strategies and physics guidance in SciFusion to isolate the contributions of unified learning, semantic factorization, and physical priors. As shown in Table 1, replacing separate single-modality models with a unified multimodal model reduces the number of models from 15 to 1 while improving generation quality. Compositional conditioning yields further gains over flattened modality embeddings, suggesting that reusable semantic factors are more effective than

Table 2: Compositional conditioning across diffusion backbones, compared with backbone-specific conditioning. Underlined values denote the better result for each backbone.

Backbone	Conditioning	VMAE		I3D		JEDI ↓
		FVD ↓	KID ($\times 10^3$) ↓	FVD ↓	KID ($\times 10^3$) ↓	
Backbone-specific baseline						
MedLoRD [21]	Flattened	28.63	106.80	11.80	21.08	2.356
Video Diffusion [10]	Original	56.87	184.97	24.00	49.28	4.63
3DMedDiff [26]	Original	91.27	373.75	33.40	78.48	8.341
Medical Diffusion [14]	Original	34.55	130.12	13.19	24.08	2.732
Compositional conditioning						
MedLoRD [21]	Compositional	27.30	102.97	11.10	19.69	2.354
Video Diffusion [10]	Compositional	51.88	168.03	23.81	48.99	4.40
3DMedDiff [26]	Compositional	26.85	97.40	11.01	18.69	2.400
Medical Diffusion [14]	Compositional	33.80	126.91	13.09	23.68	2.733
SciFusion (Ours)	Compositional	**23.51**	**84.76**	**9.90**	**16.59**	**1.797**

treating modalities as independent categories. Adding domain-specific physics guidance provides an additional improvement, achieving the best overall performance.

The benefit of compositional conditioning is not specific to our diffusion architecture: replacing backbone-specific conditioning with the proposed factorization improves both FVD and JEDI across four different diffusion backbones (Table 2), suggesting that it offers a generalizable representation strategy for heterogeneous scientific imaging modalities rather than an architecture-specific modification. A per-domain (MRI/CT/EM) breakdown is provided in the supplementary material (Table 8). Qualitative generation results are shown in Fig. 2, which presents representative generated volumes across the evaluated imaging modalities. Additional visualizations of the generated 16-frame volumes are provided in the supplementary material (Fig. 4).

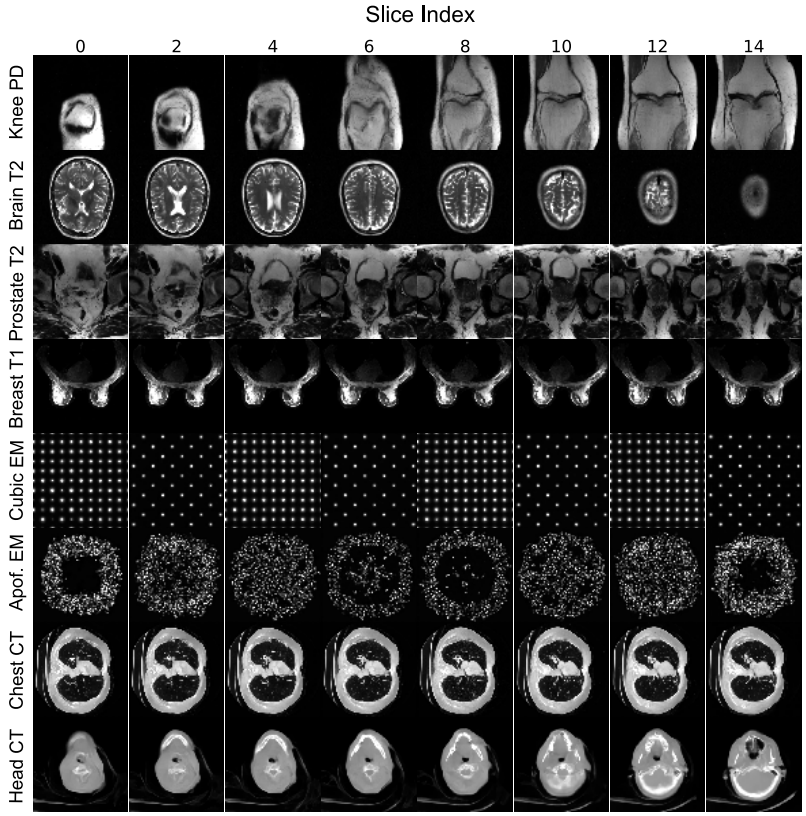

Figure 2: Representative generated 3D volumes from selected imaging modalities evaluated in this study.

3.3 Zero-shot Compositional Generation

Beyond improving generation quality on observed modalities, we next evaluate whether the learned factorized representation captures reusable structure that can be recombined for unseen modality configurations. We consider two unseen modalities, knee T1 and brain PD, for which no training samples are provided. SciFusion synthesizes each target by recombining learned domain–object–contrast representations, pairing previously observed anatomical objects with learned imaging contrasts in novel combinations without additional training.

Table 3: Zero-shot generation compared with supervised training and parameter-efficient adaptation.

Target	**Backbone**	**Supervision**	**I3D**		**VMAE**		**JEDI** $\downarrow$
			FVD $\downarrow$	**KID** $(\times 10^3)$ $\downarrow$	**FVD** $\downarrow$	**KID** $(\times 10^3)$ $\downarrow$	
Knee T1	SciFusion (Ours)	Full	**13.83**	**23.13 ± 2.00**	**14.56**	**58.71 ± 3.78**	**0.35**
	SciFusion (Ours)	Zero-shot	26.14	50.23 ± 2.39	58.32	252.47 ± 7.07	2.88
	Medical Diffusion [14]	Zero-shot	28.46	54.50 ± 2.65	72.46	322.46 ± 7.36	3.59
	VDM + LORA [10, 11]	PEFT	15.99	23.06 ± 1.39	25.63	102.29 ± 6.65	1.05
Brain PD	SciFusion (Ours)	Full	**2.90**	**4.30 ± 0.49**	**5.22**	**19.19 ± 1.25**	**0.46**
	SciFusion (Ours)	Zero-shot	23.52	63.84 ± 3.98	15.34	64.17 ± 2.36	1.68
	Medical Diffusion [14]	Zero-shot	11.02	23.54 ± 1.03	27.97	125.20 ± 4.35	2.25
	VDM + LORA [10, 11]	PEFT	14.33	30.70 ± 2.00	31.13	137.24 ± 4.07	1.75

As shown in Table 3, fully supervised models achieve the best performance due to direct access to target-modality samples, while our compositional model generates unseen modalities without any target samples, indicating that the learned representation captures transferable relationships between anatomical structures and imaging contrasts. PEFT narrows this gap on knee T1 using target-modality samples and additional optimization; notably, on brain PD our training-free zero-shot generation already matches or exceeds PEFT on JEDI (1.68 vs. 1.75), suggesting that the learned domain–object–contrast composition for this target is already well aligned without any adaptation. Qualitative results in Fig. 3 further confirm visually plausible, modality-consistent anatomical structures.

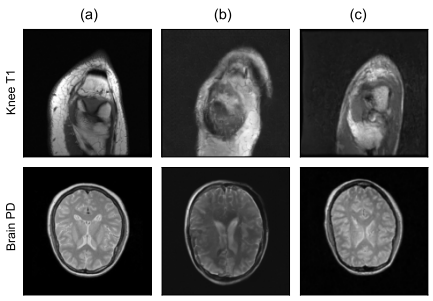

Figure 3: Qualitative comparison on target modalities: real images (a), zero-shot (b), and PEFT adaptation (c).

4 Conclusion

Extensive experiments across 15 imaging modalities show that SciFusion improves generation quality and that its compositional conditioning strategy generalizes across diffusion backbones. Experiments on two held-out MRI modality combinations further demonstrate the feasibility of zero-shot synthesis without target-domain training. SciFusion is not intended to replace fully supervised models in data-rich settings, but rather to offer a scalable alternative when target-domain data are limited. A key limitation is that zero-shot compositional transfer is currently demonstrated on only two unseen MRI combinations and may degrade under large distribution shifts; lightweight target-specific guidance or few-shot adaptation could help close this gap. Future work includes conducting broader multimodal evaluation and combining generated 3D data with scarce real-world samples across scanners, protocols, and imaging domains.

References

[1] Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Robert J. Stewart, Francis G. Blankenberg, David B. Larson, Richard H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLOS Medicine*, 15(11):e1002699, 2018.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *International Conference on Learning Representations*, 2018.

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023.

[4] Stephen K Burley, Rusham Bhatt, Charmi Bhikadiya, Chunxiao Bi, Alison Biester, Pratyoy Biswas, Sebastian Bittrich, Santiago Blaumann, Ronald Brown, Henry Chao, et al. Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. *Nucleic acids research*, 53(D1):D564–D574, 2025.

[5] John M Cowley and Alexander Forbes Moodie. The scattering of electrons by atoms and crystals. i. a new theoretical approach. *Acta Crystallographica*, 10(10):609–619, 1957.

[6] Adam E. Flanders, Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan, Anouk Stein, Felipe C. Kitamura, Matthew P. Lungren, Gagandeep Choudhary, Lesley Cala, Luiz Coelho, Monique Mogensen, Fanny Morón, Elka Miller, Ichiro Ikuta, Vahe Zohrabian, Olivia McDonnell, Christie Lincoln, Lubdha Shah, David Joyner, Amit Agarwal, Ryan K. Lee, and Jaya and Nath. Construction of a machine learning dataset through collaboration: The rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020. PMID: 33937827.

[7] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. *arXiv preprint arXiv:2305.16037*, 2023.

[8] Ji He, Yongbo Wang, and Jianhua Ma. Radon inversion via deep learning. *IEEE Transactions on Medical Imaging*, 39(6):2076–2087, 2020.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[11] 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.

[12] Information eXtraction from Images Project. IXI Dataset. `https://brain-development.org/ixi-dataset/`. Licensed under CC BY-SA 3.0; accessed August 4, 2026.

[13] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[14] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13(1):7303, 2023.

[15] Earl J Kirkland. *Advanced computing in electron microscopy*, volume 12. Springer, 1998.

[16] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. 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.

[17] I Lobato and D Van Dyck. An accurate parameterization for scattering factors, electron densities and electrostatic potentials for neutral atoms that obey all physical constraints. *Foundations of Crystallography*, 70(6):636–649, 2014.

[18] Ge Ya Luo, Gian Mario Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond FVD: Enhanced evaluation metrics for video generation quality. *ICLR*, 2025.

[19] Jacob Madsen and Toma Susi. The abtem code: transmission electron microscopy from first principles. *Open Research Europe*, 1:24, 2021.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[21] Marvin Seyfarth, Salman Ul Hassan Dar, Isabelle Ayx, Matthias Alexander Fink, Stefan O. Schoenberg, Hans-Ulrich Kauczor, and Sandy Engelhardt. Medlord: A medical low-resource diffusion model for high-resolution 3d ct image synthesis. In Virginia Fernandez, David Wiesner, Lianrui Zuo, Adrià Casamitjana, and Samuel W. Remedios, editors, *Simulation and Synthesis in Medical Imaging*, pages 1–12, Cham, 2026. Springer Nature Switzerland.

[22] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019.

[23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[24] Stanford Center for Artificial Intelligence in Medicine and Imaging. MRNet: Knee MRIs. `https://doi.org/10.71718/rcbp-8c35`. Accessed: 2026-08-04.

[25] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. *ICLR Workshop DeepGenStruct*, 2019.

[26] 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*, 44:4960–4972, 2024.

[27] Jiayi Wang, Hadrien Reynaud, Franciskus Xaverius Erick, and Bernhard Kainz. Ctflow: Video-inspired latent flow matching for 3d ct synthesis, 2025.

[28] Shuo Xu, Jintao Fu, Yuewen Sun, Peng Cong, and Xincheng Xiang. Deep radon prior: A fully unsupervised framework for sparse-view ct reconstruction. *Comput. Biol. Med.*, 189(C), May 2025.

[29] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

Supplementary Materials

A Dataset Physical Parameters

We construct a 15-modality volumetric benchmark spanning MRI, CT, and EM to evaluate unified multimodal generation and compositional generalization. MRI volumes covering knee, brain, prostate, and breast are obtained from NYU fastMRI [16, 29]. Head CT is from the RSNA 2019 Brain CT Hemorrhage Challenge [6], while chest and abdominal CT are from the Medical Segmentation Decathlon [22]. EM data include cubic and hexagonal crystals, apoferritin, and hemoglobin from the Materials Project [13] and PDB [4], converted to volumetric stacks using `abTEM` [19]. For zero-shot evaluation, knee T1 from MRNet [24, 1] and brain PD from IXI [12] are held out, while their constituent factors remain observed during training. We generate 400 samples per modality for multimodal evaluation and 600 samples for each zero-shot target. Modality-specific acquisition and preprocessing parameters are detailed below.

Magnetic Resonance Imaging (MRI). We collect volumetric data from the fastMRI dataset [29], Brain IXI [12], and Stanford Knee [24, 1]. The datasets are acquired at field strengths of 1.5 T and 3 T. All data are pre-processed to a fixed depth–height–width dimension of $16 \times 256 \times 256$; the slice dimension is chosen to preserve the most informative anatomical structures. Brain IXI and Stanford Knee are used only to assess zero-shot generation via compositional embedding and are excluded from training. Table 4 summarizes the datasets.

Table 4: Summary of MRI datasets.

Dataset	Anatomy	Contrast	Total
fastMRI	Knee	PD and PDFS	(480, 486)
fastMRI	Brain	T1, T2, FLAIR	(328, 1367, 2547)
fastMRI	Prostate	T2 and DWI	(272, 261)
fastMRI	Breast	T1	609
IXI	Brain	PD	566
MRNet	Knee	T1	1130

Computed Tomography (CT). We collect CT data from the RSNA Brain dataset [6] and the Medical Segmentation Decathlon [22]. Images are pre-processed to 256×256 with 16 slices. Table 5 summarizes the datasets.

Table 5: Summary of CT datasets.

Dataset	Anatomy	Total
RSNA	Brain	339 exams (subset)
MSD	Liver	201 scans
MSD	Lung	96 scans
MSD	Pancreas	420 scans
MSD	Colon	190 scans
MSD	Liver	443 scans
MSD	Spleen	61 scans

Electron Microscopy (EM). Table 6 summarizes the abTEM simulation parameters for crystalline and protein datasets, including cell size, spatial resolution, slice discretization, and propagation settings; both use plane-wave illumination. For crystal materials, we construct cubic and hexagonal structures with 278 and 160 volumes, respectively. For proteins, we collect 69 apoferritin and 204 hemoglobin volumes.

Table 6: abTEM simulation parameters used for dataset generation.

Parameter	Crystal data	Protein data
Pixel dimension	256×256	256×256
Sampling / pixel size	0.1 Å	1.0 Å
Number of slices (S)	16	16
Slice thickness	0.5 Å	4.0 Å
Potential	Lobato [17]	Lobato [17]
Incident wave	Plane wave	Plane wave
Electron energy	300 keV	300 keV

B Implementation Details

Computing Resources. In our experiments, we leveraged high-performance computing (HPC) nodes equipped with 8 NVIDIA A100 GPUs with 80GB VRAM each and 128 CPU cores.

VQGAN Training Details. We employ a 3D VQGAN as the first-stage autoencoder to learn a compact latent representation for high-resolution scientific imaging volumes. The VQGAN is trained on the complete multi-modal dataset, spanning all imaging domains and modalities, without any modality-specific supervision or target-modality adaptation.

The input volume size is $16 \times 256 \times 256$. The encoder compresses each volume into a latent representation with 8 channels and spatial resolution $16 \times 32 \times 32$, subsequently used by the latent diffusion model. The vector-quantization module uses a codebook of 16,384 embeddings with embedding dimension 8 and commitment cost 0.25.

The VQGAN is trained for 1,000 epochs with batch size 1, using Adam with learning rates 3×10^{-5} (generator) and 2×10^{-5} (discriminator). The training objective combines reconstruction, perceptual, adversarial, and 3D feature-based Jukebox losses, weighted at 0.1, 0.008, and 0.7 for the perceptual, adversarial, and Jukebox terms, respectively; adversarial training begins after a 10-epoch warm-up. The discriminator is a two-layer 3D convolutional network with 64 channels and instance normalization. The VQGAN uses three encoder–decoder stages with channel dimensions $[128, 256, 512]$, two residual layers per stage, and ReLU activations.

Latent Diffusion Training Details. The diffusion model operates on precomputed VQGAN latents (8 channels, $16 \times 32 \times 32$) rather than directly in image space, using a 3D U-Net with channel dimensions $[64, 128, 256, 512]$ and two residual blocks per resolution level. Attention modules are applied at the two highest-resolution stages; cross-attention conditioning uses a feature dimension of 64 with attention heads of 32 channels.

The compositional conditioning representation spans 3 domain, 12 object, and 9 contrast categories, each embedded in a 64-dimensional space and combined as described in Sec. 2.2, enabling generation of unseen domain–object–contrast combinations without additional target-modality training.

The diffusion model is trained for 500 epochs with batch size 4 using Adam, learning rate 1×10^{-4} decayed by a factor of 0.9995, and an EMA decay of 0.999 for optimization stability. We use a cosine noise schedule with 300 diffusion timesteps and the velocity (v)-prediction parameterization, trained with a unified objective across all domains and modalities. The physics-guided regularization term (Sec. 2.3) is introduced via a warm-up strategy, activated at epoch 100 and increased gradually until epoch 150, with a base physical-loss scale of 0.05; this allows the model to first learn the general latent distribution before introducing physical constraints.

C Electron Microscopy Wave Propagation

Section 2.3 defines the EM physical guidance operator $\mathcal{P}_{\mathrm{EM}}$ from the far-field diffraction intensity after multislice wave propagation. Here we detail the propagation recursion. Given plane-wave illumination ψ_0, the wave function after the n-th slice is computed as

$$\psi_n = \mathcal{F}_{2D}^{-1}[\mathcal{F}_{2D}(\psi_{n-1} \exp(iV_n))\, P]\,, \tag{2}$$

where V_n denotes the n-th slice of the projected electrostatic potential, $\mathcal{F}_{2D}$ and $\mathcal{F}_{2D}^{-1}$ are the two-dimensional Fourier and inverse Fourier transforms, and P is the Fresnel propagation kernel. Each slice thus modulates the incoming wave by its local potential and propagates it to the next slice via Fresnel diffraction. After propagation through all S slices, the far-field diffraction representation used in $\mathcal{P}_{\mathrm{EM}}$ is obtained from the intensity of the final wave function ψ_S. Simulation parameters (pixel size, slice thickness, electron energy) are given in Table 6.

D Physical Loss Weight Sensitivity

As noted in Sec. 2.3, we select the physical-loss weight λ_{phys} in the overall training objective $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$ via a sweep over $\lambda_{\mathrm{phys}} \in \{2, 4, 5\} \times 10^{-4}$, training separate models under the same configuration for each value and evaluating them at a matched training stage across MRI, CT, and EM domains. Table 7 reports the full per-domain and overall results underlying our choice of $\lambda_{\mathrm{phys}} = 5 \times 10^{-4}$.

Table 7: Effect of physical-loss regularization strength across imaging domains. KID values are multiplied by 10^3 for readability.

Domain	$\boldsymbol{\lambda_{\mathrm{phys}}}$	**I3D**		**VMAE**		**JEDI** $\downarrow$
		FVD $\downarrow$	**KID** $\downarrow$	**FVD** $\downarrow$	**KID** $\downarrow$	
MRI	$5 \cdot 10^{-4}$	**10.18**	**18.04**	**24.70**	**95.35**	**1.46**
	$4 \cdot 10^{-4}$	11.59	22.26	28.69	114.18	1.64
	$2 \cdot 10^{-4}$	11.46	21.40	28.95	115.29	1.64
EM	$5 \cdot 10^{-4}$	5.42	6.28	16.85	46.68	**1.50**
	$4 \cdot 10^{-4}$	**5.04**	**5.59**	**15.15**	**38.46**	2.13
	$2 \cdot 10^{-4}$	5.87	7.51	18.06	50.54	2.44
CT	$5 \cdot 10^{-4}$	**16.72**	**30.38**	**31.81**	**123.55**	**3.08**
	$4 \cdot 10^{-4}$	21.89	45.18	36.66	142.38	3.37
	$2 \cdot 10^{-4}$	17.83	33.58	34.92	136.25	3.45
Overall	$5 \cdot 10^{-4}$	**9.90**	**16.59**	**23.51**	**84.77**	**1.80**
	$4 \cdot 10^{-4}$	11.47	21.29	25.77	94.58	2.15
	$2 \cdot 10^{-4}$	10.87	19.20	26.51	97.90	2.27

As shown in Table 7, the optimal weight varies slightly across domains – for instance, EM achieves marginally better FVD and KID with $\lambda_{\mathrm{phys}} = 4 \times 10^{-4}$ – reflecting the distinct physical properties and dynamic ranges of each domain's guidance operator ($\mathcal{P}_{\mathrm{MRI}}$, $\mathcal{P}_{\mathrm{CT}}$, $\mathcal{P}_{\mathrm{EM}}$). However, when averaged across all three domains, $\lambda_{\mathrm{phys}} = 5 \times 10^{-4}$ achieves the best overall performance across all metrics and is adopted as the default value for all experiments reported in the main paper.

E Additional Qualitative Results

Fig. 4 visualizes generated volumes across all 15 imaging modalities, complementing the representative subset shown in the main paper.

F Additional Quantitative Results

Table 8 reports a domain-level breakdown of the backbone-generalization comparison in the main paper (Sec. 2), comparing backbone-specific baseline conditioning against the proposed compositional conditioning separately for each of the three imaging domains (MRI, EM, CT), macro-averaged over eight MRI, four EM, and three CT modalities.

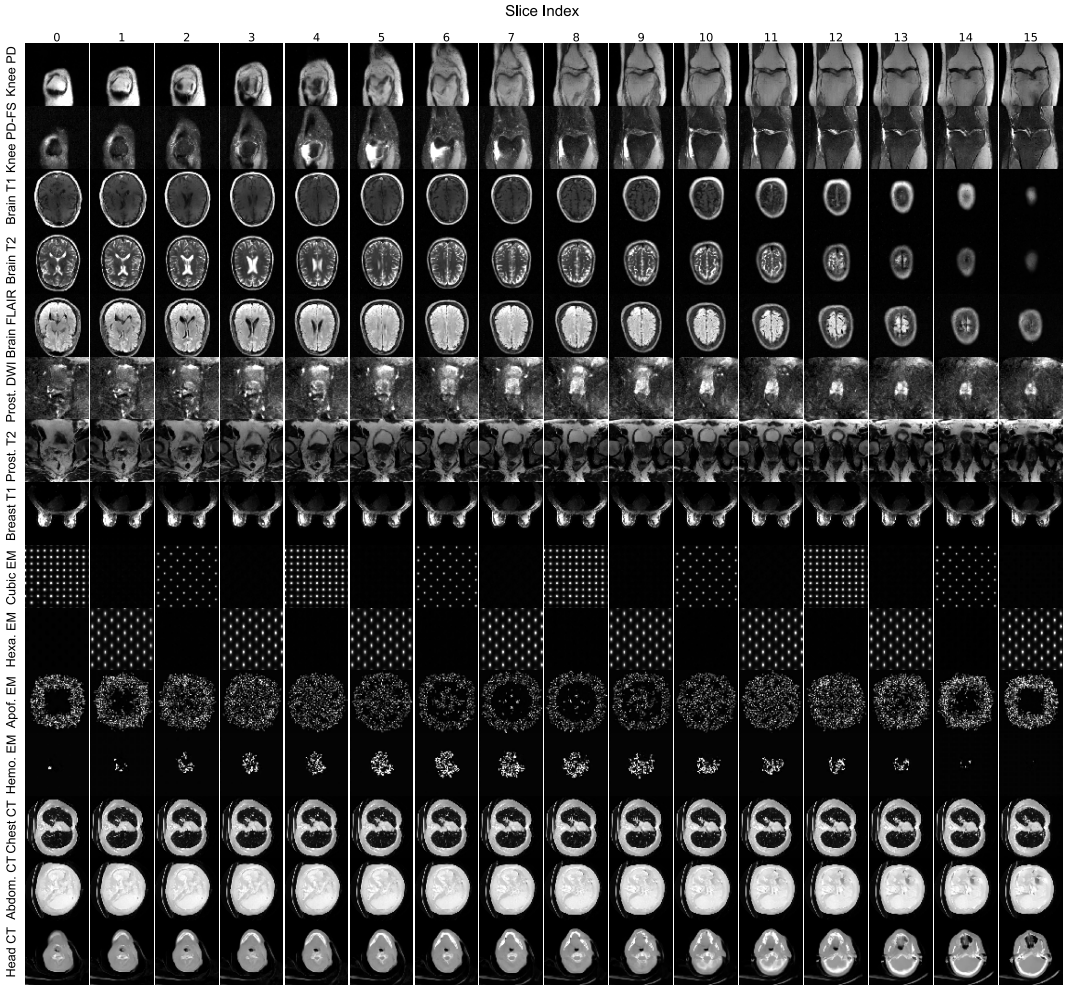

Figure 4: Representative generated 3D volumes across all imaging modalities. Each row corresponds to one modality, and the 16 columns visualize axial slices from a single generated volume.

Table 8: Domain-level comparison between backbone-specific baseline conditioning and the proposed compositional conditioning across three imaging domains. Results are macro-averaged over modalities within each domain, including eight MRI modalities, four EM modalities, and three CT modalities. For MedLoRD, the backbone-specific baseline uses an independently learned embedding for each modality, whereas compositional conditioning uses the proposed domain–object–contrast representation. Other backbones retain their original conditioning mechanisms for the baseline setting.

Backbone	**Domain**	**VideoMAE-FD** ↓	**VideoMAE-KID** ↓ $\times 10^3$	**I3D-FD** ↓	**I3D-KID** ↓ $\times 10^3$	**JEDI** ↓
Backbone-specific baseline conditioning						
MedLoRD	MRI	27.33	109.08	11.21	20.10	1.668
	EM	17.38	43.82	3.93	3.13	2.522
	CT	47.13	184.68	23.85	47.63	3.969
Pixel Diffusion	MRI	45.84	141.80	29.28	63.99	2.042
	EM	87.48	297.31	9.81	15.93	10.519
	CT	45.48	150.33	28.85	54.51	3.66
3DMedDiff	MRI	107.39	459.01	42.07	102.67	8.092
	EM	35.74	110.34	5.94	7.27	5.480
	CT	122.30	497.63	46.92	108.90	12.816
Medical Diffusion	MRI	36.97	150.29	14.30	27.09	1.694
	EM	22.64	59.03	5.34	7.68	3.769
	CT	43.99	171.12	20.68	37.90	4.120
Compositional conditioning						
MedLoRD	MRI	27.45	109.88	11.65	21.76	1.718
	EM	18.49	51.36	4.14	4.09	2.662
	CT	38.67	153.33	18.88	35.00	3.642
Pixel Diffusion	MRI	40.41	125.66	29.36	64.10	1.949
	EM	84.43	282.65	9.89	16.43	9.889
	CT	39.07	128.17	27.56	52.13	3.64
3DMedDiff	MRI	27.97	110.11	12.17	22.46	1.379
	EM	22.49	64.49	4.06	3.73	3.593
	CT	29.65	107.37	17.16	28.55	3.530
Medical Diffusion	MRI	35.86	146.07	14.26	27.16	1.690
	EM	22.50	58.82	4.97	6.32	3.644
	CT	43.35	166.60	20.81	37.55	4.299
SciFusion (Ours)	MRI	**22.95**	**87.56**	**9.99**	**17.11**	**1.43**
	EM	**18.40**	**50.11**	**4.63**	**5.19**	**1.57**
	CT	**31.81**	**123.55**	**16.72**	**30.38**	**3.08**

Note: Underline values indicate the better result between baseline and compositional conditioning for the same backbone and domain.