Unified Multimodal Model for Brain MRI Imputation and Understanding

Zhiyun Song[1✉], Che Liu[1✉], Tian Xia[1], Avinash Kori[1], and Wenjia Bai[1,2]

[1] Department of Computing, Imperial College London, London, UK
[2] Department of Brain Sciences, Imperial College London, London, UK
z.song25, che.liu21@imperial.ac.uk

Abstract. Multimodal large language models (MLLMs) hold great potential for medicine, as they inherit knowledge from LLM and allow multiple data modalities to be integrated, analysed and interpreted in natural language. However, the field of medical MLLMs is constrained by non-trivial challenges, notably the scarcity of high-quality training data and the frequent occurrence of missing data in the real-world clinical setting. Here, we propose a novel unified multimodal model, UniBrain, for brain magnetic resonance image (MRI) analysis. To address potential missing brain MRI modalities, we employ a unified training strategy to perform joint imaging modality imputation and brain image understanding. During training, an interleaved and description-enriched data flow is constructed to train the model in an autoregressive manner, enabling medical reasoning with generated multimodal data. A self-alignment strategy is introduced to leverage dense image embeddings to learn fine-grained anatomical features without requiring detailed image captions. Furthermore, we propose a dynamic hidden state mechanism to alleviate the exposure bias during long-context multimodal inference. Extensive experiments on multi-disease brain MRI dataset demonstrate that UniBrain achieves high performance for brain image imputation, understanding, and disease diagnosis under various extents of modality incompleteness. Source code is publicly available at https://github.com/zhiyuns/UniBrain.

Keywords: Unified multimodal model · Brain MRI · Brain image analysis · Modality imputation.

1 Introduction

Brain magnetic resonance imaging (MRI) plays an indispensable role in identifying structural abnormalities of the brain, diagnosing, and monitoring the progression of various neurological diseases. Recently, multimodal large language models (MLLMs) have demonstrated potential in leveraging knowledge learnt from LLMs combined with modality-specific encoders to integrate and interpret multimodal data, including medical images, within a single model [1,2,3,30]. However, developing MLLMs to interpret brain MRI scans faces significant challenges. Reading brain MRI scans requires radiological and neurological expertise,

which may not be well captured by a general-purpose MLLM [2]. Moreover, brain MRI can consist of multiple modalities, *e.g.* T1-weighted, T2-weighted, FLAIR etc. Disease diagnosis would ideally require a comprehensive view of all modalities. In clinical reality, however, some modalities may be missing [23], which creates the challenge for the MLLM to deal with the data modality gap.

To deal with missing modalities, existing literature generally follows two paradigms: *explicit modality imputation*, and *implicit representation learning* [27]. Explicit methods leverage the available modalities to synthesise images of missing modalities, providing input for downstream tasks [4,20,24]. While these methods prioritise the fidelity of the synthesised images, imputation and subsequent analysis operate as a disjointed, two-stage pipeline, which might compromise the performance in learning downstream task-relevant representations from synthesised images. In contrast, implicit methods deal with missing modalities implicitly in the model. Some of these methods align the representations of the available and missing modalities within a shared space [18,32]. Some others employ simple generative modules to map representations of available images to those of the missing ones [7,31]. The objective of implicit representation learning is to improve the performance of downstream tasks. It may sacrifice the fidelity of the synthesised images. Overall, both paradigms require further exploration of the mutual benefits between generation and understanding.

The recent emergence of unified multimodal models for generation and understanding offers a promising solution [5,12,26]. By tokenising and modeling both visual and textual data within a single, autoregressive decoder-only architecture, the unified models natively intertwine the two tasks of generation and understanding. This enables multimodal interleaved reasoning, such as thinking with generated images [17], where the generation process enhances the chain-of-thought reasoning. However, developing unified models for medical domain faces three critical obstacles. First, there is a lack of interleaved and multimodal medical dataset for training a unified model. How to design tasks that effectively leverage the generation ability for multimodal understanding is still under-explored. Furthermore, there is a domain gap between natural and medical image-text pairs. Models trained on natural images are ill-equipped with specialist knowledge to understand medical images and learn clinically meaningful representations [15]. Finally, interleaved reasoning relies on intermediate generation results as a context to perform subsequent reasoning, which may lead to error accumulation, compromising the final model performance [21].

In this work, we propose UniBrain, a novel multimodal model to perform brain image analysis with missing MRI modalities. As a unified model, UniBrain simultaneously improves both generation fidelity and diagnostic accuracy through a multimodal autoregressive process. To overcome the aforementioned obstacles, we introduce three core contributions: 1) We design a novel framework for brain MRI diagnosis, which imputes missing MRI modalities as intermediate steps before formulating the final understanding output, enabling reasoning with imputed images. 2) We introduce a self-adaption strategy that leverages the visual embeddings of the model to reconstruct input medical images, which

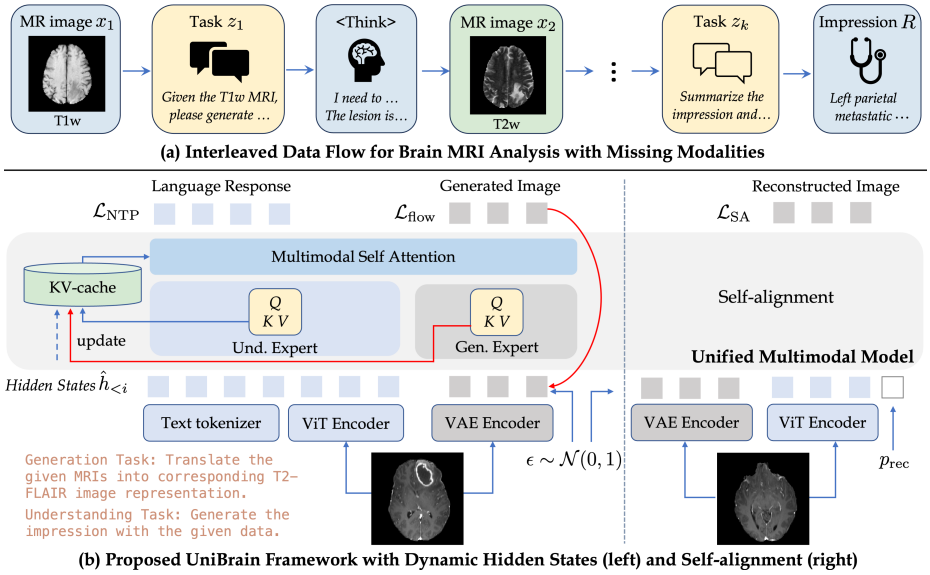

Fig. 1. Overall framework of UniBrain. (a) We construct a sequential, interleaved data flow for brain MRI analysis during training. (b) The framework incorporates a unified multimodal model for image generation and understanding. Dynamic hidden states via training-time KV-cache are designed to mitigate exposure bias for long-context multimodal reasoning. A self-alignment loss conditions the generation expert on dense embeddings extracted from the ViT for image reconstruction.

reduces the need for medical reports with fine-grained visual descriptions. 3) To mitigate accumulated errors in long-context autoregressive generation, we implement a dynamic hidden state mechanism that enforces the model to be aware of the self-generated artefacts during training, improving the robustness of autoregressive inference.

2 Methodology

2.1 Preliminary of Unified Multimodal Model

UniBrain is built upon BAGEL [5], a unified model for joint vision-language understanding and generation. Employing a Mixture-of-Transformer-Experts (MoT) architecture [11], BAGEL comprises two Transformer experts, respectively dedicated to understanding and generation. Given multimodal input, text is encoded by a text tokeniser, whereas images are encoded into visual tokens by two encoders, an understanding-oriented ViT and a generation-oriented VAE. The two experts operate over shared text and visual tokens via multimodal self-attention, which allows long-range interactions between the tokens. On the output side, BAGEL predicts the next token for language response and visual tokens for

image generation via rectified flow [5]. UniBrain adapts BAGEL to brain MRI analysis in three aspects, as explained in the following sections.

2.2 Unified Brain MRI Modality Imputation and Understanding

We formulate brain MRI missing modality imputation and image understanding (*e.g.* disease diagnosis) as an autoregressive sequence modeling problem. We denote the images of k MRI modalities as $X=\{x_i\}_{i=1}^{k}$, the corresponding radiological findings for each modality as $F=\{f_i\}_{i=1}^{k}$, and the impression as R which contains global description and diagnostic results. As illustrated in Fig. 1(a), given input image x_1, a multimodal model π_θ, parameterised by θ, iteratively executes reasoning-enriched generation tasks $Z_g=\{z_i\}_{i=1}^{k-1}$ and finally performs the understanding task $Z_u=z_k$. The overall interleaved data flow is constructed as $\{x_1, z_1, f_1, x_2, \ldots, z_k, R\}$. The model contains hidden state h_i, which are tokenised images or text. It is trained to predict its next state based on previous accumulated multimodal context: $\hat{h}_i=\pi_\theta(x_1, z_i, h_{<i})$, where $h_{<i}=\{h_1, h_2, \ldots, h_{i-1}\}$ denotes the history of hidden states. For generation task $z_i \in Z_g$, the model is instructed to perform the thinking procedure before generating the target image x_{i+1}. This thinking procedure consists of two components: 1) enriched description of the current objective z_i (*e.g.*, "*I need to translate the given T1 and FLAIR images into the corresponding T2-weighted MRI*".), and 2) the radiological findings ($f_{\leq i}$) associated with the already known (ground-truth or synthesised) input modalities. For the final understanding task z_k, the model leverages the full accumulated sequence to produce the overall impression R.

2.3 Fine-grained MRI Representation with Self-alignment

One challenge in adapting general-purpose unified models to medical imaging is to bridge the gap between medical imaging modalities and free-form texts. Conventional vision-language alignment relies heavily on abundant paired natural image-text data, while paired medical image-text data are extremely sparse. In addition, radiological text such as "*hyperintense lesion in the right parieto-occipital lobe with ill-defined margins*" only provides a high-level description of a brain MRI scan and does not provide pixel-level supervisory signals. This leads to the discrepancy between the ViT tokens used for high-level understanding and VAE tokens used for pixel-level MRI generation.

To overcome the limitation, we propose a self-alignment (SA) strategy that leverages intrinsic supervisory signals from the images themselves. Drawing inspiration from [28], we postulate that visual embeddings extracted from a model's visual understanding encoder may contain richer semantic representations of an image than what radiological text caption could capture. Using these visual embeddings as prompts, a unified model can be trained to reconstruct the original image. In detail, we condition the generation expert on the dense visual embeddings of image x_i extracted from the ViT encoder. As the proposed model performs visual token prediction in the latent space, the self-supervised image reconstruction objective is formulated accordingly. The VAE encoder projects

x_i into a latent representation $e_i = E_{\text{VAE}}(x_i)$, while the ViT encoder extracts the dense visual embeddings $w_i = E_{\text{ViT}}(x_i)$. The training objective for self-alignment is to predict the velocity in the latent space of the target MRI modality, conditioned on the visual embeddings w_i and the text prompt p_{rec} to instruct reconstruction. The self-alignment loss is formulated as:

$$\mathcal{L}_{\text{SA}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\epsilon\sim\mathcal{N}(0,I)}[||V_\theta(e_i^t, t, w_i, p_{\text{rec}}) - (\epsilon - e_i)||^2], \tag{1}$$

where $e_i^t = (1-t)e_i + t\epsilon$ represents the interpolated latent between the clean latent e_i and the noise sample ϵ. V is the velocity prediction network [6] parameterised by the generation expert. As the training of $\mathcal{L}_{\text{SA}}$ requires only the image itself, the model learns from both highly pathological images and healthy images, alleviating the need for detailed visual descriptions of these images.

2.4 Bridging the Training-test Gap via Dynamic Hidden States

Another challenge for unified models is the training-test domain gap, commonly referred to as the exposure bias [19]. This gap originated from the fact that autoregressive inference is based on the previous predicted states $\hat{h}_{<i}$, which are generated and thus with a bias from the ground-truth states $h_{<i}$ (used during training). Although BAGEL implements a diffusion forcing strategy to alleviate the exposure bias, it operates in the ground-truth noisy latent space, leading to accumulated errors for the context $(x_1, z_i, \hat{h}_{<i})$ when i increases.

To bridge this gap, we propose a dynamic hidden state (DHS) mechanism that enforces the model to condition on its own visual predictions during training. Due to the long-context scenario that demands massive computation, we leverage a training-time KV cache mechanism [8] to enable an affordable end-to-end autoregressive rollout of the constructed data flow. We first process the deterministic, known information. The system instructions and input modality x_1 are encoded and tokenised, with their activations being stored in the KV cache. When the data flow reaches a generation token (*i.e.*, tokenised e_i^t), we initiate a few-step (10 steps in our case) diffusion [8] in the VAE latent space to generate the current missing modality, where generation loss is computed on the exit timestep. This generation is conditioned on the previously KV cache containing history hidden states. Then, we project the self-generated visual tokens $\hat{e}_i$ back to the model's hidden state $\hat{h}_i$ and append it to the KV cache. Finally, the understanding task z_k is processed using the fully accumulated KV cache. Because this cache now contains the self-generated dynamic intermediate states rather than the ground-truth static ones, the model is forced to provide accurate results despite the presence of its own generative artefacts.

Following the procedure, UniBrain is trained with a unified loss function modified with dynamic states. For text-related tasks (thinking and reporting), we employ the next-token prediction (NTP) loss for each hidden state:

$$\mathcal{L}_{\text{NTP}} = -\sum_n \log p(h_{i,n}|\mathbf{KV}_{<i}, h_{i,<n}; \theta), \tag{2}$$

where $h_{i,n}$ denotes the n-th token in its current state, $\mathbf{KV}_{<i}$ denotes the KV cache for previous states. For visual generation, we apply flow matching on the latents:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t\sim U[0,1],\epsilon\sim\mathcal{N}(0,I)}[||V_\theta(e_i^t, t, \mathbf{KV}_{<i}) - (\epsilon - e_i)||^2], \tag{3}$$

where the model predicts the velocity for interpolated latent e_i^t conditioned on the previous KV cache. We keep the bidirectional attention on visual tokens in current states. The overall loss can be formulated as:

$$\mathcal{L} = \frac{1}{k}\sum_i (\mathcal{L}_{\text{SA}} + \mathcal{L}_{\text{NTP}} + \mathcal{L}_{\text{flow}}). \tag{4}$$

During inference, each patient case provides a available modalities $\{x_i\}_{i=1}^{a}$ as inputs, while the remaining modalities $\{x_i\}_{i=a+1}^{k}$ are missing and need to be synthesised. We sequentially perform the generation tasks $\{z_i\}_{i=a}^{k-1}$, before performing the final understanding task z_k. For each task, the model predicts its next state based on the accumulated multimodal context: $\hat{h}_i{=}\pi_\theta(x_1, z_i, \hat{h}_{<i})$, where $\hat{h}_{<i}{=}\{h_1, \ldots, h_a, \hat{h}_{a+1}, \ldots, \hat{h}_{i-1}\}$ denotes the history of hidden states.

3 Experiments

3.1 Data and Experiment Settings

Dataset A public dataset, RadGenome-Brain MRI [9], is used for model training and evaluation, which contains 3,408 brain MRI scans acquired from multiple sites with paired radiology reports, covering six MRI modalities (T1, T2, FLAIR, DWI, ADC, and T1ce) and five neurological disease types (glioma, meningioma, metastasis, stroke, white matter hyperintensities (WMH)). The reports were written by senior radiologists, providing findings and impressions grounded in the annotated image regions [9]. We follow the official setting to split the dataset into training (n=2,382), validation (n=338), and test sets (n=688).

Implementation Details UniBrain is built upon the BAGEL backbone with 14B parameters pretrained on large-scale multimodal data [5]. To stabilise training, the model was trained with self-alignment for 2,000 steps, followed by 2,000 steps for non-KV cache training and 200 steps using the DHS mechanism. At each step, we sample a batch of brain MRI slices corresponding to annotated lesions and randomise the number and order of MRI modalities. The corresponding impressions and clinical findings, together with the images are packed into a sequence of 13,408 tokens for model training. Since 2D slices are used, we reformulate the 3D volumetric and size-related descriptions in the original reports and project them to 2D slice-specific descriptors. We randomly drop the text and visual tokens with the probability of 0.1 and 0.3, respectively. All experiments were performed on 8 Nvidia A100 GPUs.

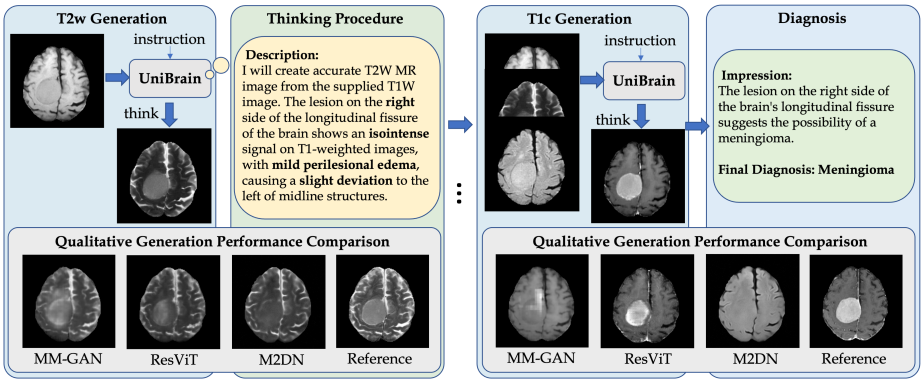

Fig. 2. Qualitative results for unified brain MRI analysis using UniBrain.

Table 1. Quantitative results for diagnosis and report generation with missing MRIs. Und. means the understanding-only counterparts of the model.

Methods	T1w only			T1w + T2w			T1w + T2w + T2f			Complete Data		
	Top-1	ROUGE	RaTE	Top-1	ROUGE	RaTE	Top-1	ROUGE	RaTE	Top-1	ROUGE	RaTE
Implicit Representation Learning												
ShaSpec [25]	63.61	-	-	71.98	-	-	75.26	-	-	77.19	-	-
SimMLM [10]	65.98	-	-	74.47	-	-	76.60	-	-	78.72	-	-
Explicit Modality Imputation												
ResViT [4] + Lingshu [29]	24.82	20.79	37.83	31.21	19.06	34.83	37.59	19.41	37.08	-	-	-
ResViT [4] + UniBrain	59.57	35.10	56.62	67.38	36.49	59.18	73.05	38.28	59.88	-	-	-
M2DN [14] + UniBrain	56.03	33.93	57.76	75.18	36.38	60.08	76.60	38.36	60.98	-	-	-
Multimodal Large Language Models												
BAGEL [5]	11.35	17.02	44.37	17.02	15.43	41.08	22.70	16.12	42.35	28.37	17.76	43.19
UniMedVL [16]	29.79	13.90	42.58	30.50	14.73	41.84	32.62	13.81	43.58	38.30	13.70	43.97
Lingshu [29]	21.99	18.08	36.91	24.82	19.49	35.55	29.79	18.13	34.32	41.13	20.26	37.29
UniBrain Und.	69.50	**37.35**	61.50	73.05	35.94	60.48	76.60	38.05	61.04	**82.06**	**38.94**	**61.62**
UniBrain	**74.47**	36.93	**61.57**	**76.60**	**38.23**	**61.24**	**78.01**	**38.68**	**61.90**	**82.06**	**38.94**	**61.62**

3.2 Results

Diagnosis and Report Generation Performance with Missing Modalities Table 1 reports the top-1 diagnostic accuracy (Top-1) [2] and report generation quality (ROUGE-1 [13], RaTEScore [33]) under different missing-modality scenarios, compared between different methods that handle missing modalities. While implicit representation learning methods (ShaSpec[25] and SimMLM [10]) achieve reasonable diagnostic accuracy, they lack clinical interpretability. Explicit modality imputation methods (ResViT [4] and M2DN [14]) perform poorly in diagnostic accuracy. This is because the imputation methods are trained separately from the MLLMs, leading to misaligned features between imputed images and MLLMs. Instead, UniBrain overcomes this issue via unified modeling and outperforms other MLLMs. It achieves 74.47% diagnostic accuracy in the highly challenging setting with only T1w modality available. With complete data, the diagnostic accuracy further increases to 82.06%.

Table 2. Quantitative evaluation of missing modality imputation. * means we ensemble results sampled with 5 random seeds.

Methods	Architecture	T1w → T2w			T1w, T2w → T2f			T1w, T2w, T2f → T1c		
		PSNR	SSIM	Top-1	PSNR	SSIM	Top-1	PSNR	SSIM	Top-1
Generative Baselines										
MM-GAN [22]	GAN	23.08	0.8774	56.74	23.32	0.8719	56.03	23.40	0.8702	60.19
ResViT [4]	GAN	22.81	0.8751	57.45	23.13	0.8740	67.38	23.00	**0.8735**	61.70
M2DN [14]	Diffusion	22.79	0.8019	51.06	22.46	0.7764	51.06	22.05	0.7820	61.70
Unified Multimodal Models										
UniMedVL [16]	Unified Model	19.82	0.7230	56.03	19.96	0.6608	63.12	21.53	0.6599	56.74
UniBrain	**Unified Model**	22.23	0.8637	**68.09**	22.58	0.8613	67.38	22.26	0.8617	74.47
UniBrain*	**Unified Model**	**23.43**	**0.8850**	63.83	**23.49**	**0.8760**	**68.08**	**23.52**	0.8725	**76.60**

Table 3. Ablation studies for UniBrain.

Model	Components			Understanding (T1w only)			Generation (T1w→ ... →T1c)		
	Unified	SA	DHS	Acc-1	ROUGE	RaTEScore	PSNR	SSIM	Top-1
A				70.05	35.71	60.12	-	-	-
B	✓			**75.11**	36.35	60.52	21.28	0.8329	70.12
C	✓	✓		73.76	35.34	59.23	22.09	0.8456	74.03
UniBrain	✓	✓	✓	74.47	**36.93**	**61.57**	**22.47**	**0.8519**	**76.60**

Missing Modality Imputation Fidelity and Diagnostic Usability Table 2 reports the fidelity of the generated missing modalities in terms of peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and the usability for downstream diagnostic task in terms of Acc-1, evaluated via the understanding expert of UniBrain. Compared to a medical multi-modal model UniMedVL [16], UniBrain achieves better performance in both imputation fidelity and diagnostic usability. Compared to strong generative model baselines, which are image-only models specifically designed for imputation, UniBrain achieves similar or slightly lower performance in imputation fidelity. However, it achieves much higher diagnostic usability in terms of Acc-1. Also, we notice that if UniBrain is ensembled with five random seeds, it can outperform specifically-designed imputation models in terms of low-level fidelity metrics (PSNR and SSIM) while maintaining superior diagnostic performance.

Ablation Studies Table 3 reports quantitative results for ablation studies to evaluate the contributions of each component in UniBrain. We start with a vanilla BAGEL architecture (A) and perform the understanding task, serving as a strong baseline. Then, we introduce interleaved data for unifed modeling (B) and enable understanding with imputed modalities, which greatly improve the diagnosis performance (by 5.06%). Furthermore, we impose self-alignment (SA) strategy (C) to enrich fine-grained representation, which benefits the generation quality (+0.81 in PSNR) while slightly disturbing the diagnosis accuracy. The final UniBrain model introduces the dynamic hidden state (DHS) mechanism, achieving optimal overall performance for both generation and understanding.

4 Conclusion

In this paper, we introduce UniBrain, a novel unified multimodal model designed to perform multimodal brain MRI analysis with potential missing modalities. UniBrain adapts an autoregressive architecture to represent fine-grained MRI details via self-alignment and conduct multimodal reasoning with the dynamic hidden state mechanism. Extensive evaluations demonstrate that UniBrain achieves strong performance in terms of modality imputation fidelity, report generation quality, and disease diagnosis accuracy. Future work will explore improving computational efficiency (currently requiring over 32GB GPU memory for inference), involving radiologists in model evaluation for diverse disease cohorts, adapting the model from 2D to 3D, and extending data modalities to further enrich the diagnostic context and advance towards multimodal AI for healthcare.

Acknowledgments. This work was supported by Imperial College London President's PhD Scholarship, EPSRC CVD-Net Programme Grant (EP/Z531297/1) and BHF New Horizons Grant (NH/F/23/70013). A.K. was supported by the EPSRC Doctoral Prize Fellowship. T.X. is supported through the Imperial College London UKRI Impact Acceleration Account EP/X52556X/1.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M.A., Damseh, R., Sheikh, J.: Multimodal large language models in health care: applications, challenges, and future outlook. Journal of Medical Internet Research **26**, e59505 (2024)
2. Bercea, C.I., Li, J., Raffler, P., Riedel, E.O., Schmitzer, L., et al.: NOVA: A benchmark for rare anomaly localization and clinical reasoning in brain MRI. In: Neural Information Processing Systems (2025)
3. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., et al.: Towards injecting medical visual knowledge into multimodal LLMs at scale. In: Conference on Empirical Methods in Natural Language Processing (2024)
4. Dalmaz, O., Yurt, M., Çukur, T.: ResViT: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10) (2022)
5. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning (2024)
7. Guo, Z., Jin, T., Zhao, Z.: Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Proceedings of the Association for Computational Linguistics (2024)
8. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: Neural Information Processing Systems (2025)

9. Lei, J., Zhang, X., Wu, C., Dai, L., Zhang, Y., et al.: Interpretable brain MRI report generation anchored by lesion topography. IEEE Journal of Biomedical and Health Informatics (2025)
10. Li, S., Chen, C., Han, J.: SimMLM: A simple framework for multi-modal learning with missing modality. In: International Conference on Computer Vision (2025)
11. Liang, W., Yu, L., Luo, L., Iyer, S., Dong, N., et al.: Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. Transactions on Machine Learning Research (2025)
12. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., et al.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL Workshop on Text Summarization Branches Out (2004)
14. Meng, X., Sun, K., Xu, J., He, X., Shen, D.: Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. IEEE Transactions on Medical Imaging **43**(7), 2587–2598 (2024)
15. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition (2025)
16. Ning, J., Li, W., Tang, C., Lin, J., Ma, C., et al.: UniMedVL: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. arXiv preprint arXiv:2510.15710 (2025)
17. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., et al.: Uni-CoT: Towards unified chain-of-thought reasoning across text and vision. In: International Conference on Learning Representations (2026)
18. Rahimpour, M., Bertels, J., Radwan, A., Vandermeulen, H., Sunaert, S., et al.: Cross-modal distillation to improve MRI-based brain tumor segmentation with missing MRI sequences. IEEE Transactions on Biomedical Engineering **69**(7) (2021)
19. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. In: International Conference on Learning Representations (2016)
20. Rassmann, S., Kügler, D., Ewert, C., Reuter, M.: Regression is all you need for medical image translation. IEEE Transactions on Medical Imaging (2026)
21. Schmidt, F.: Generalization in generation: A closer look at exposure bias. In: Proceedings of the 3rd Workshop on Neural Generation and Translation (2019)
22. Sharma, A., Hamarneh, G.: Missing MRI pulse sequence synthesis using multimodal generative adversarial network. IEEE Transactions on Medical Imaging (2020)
23. Shen, Y., Gao, M.: Brain tumor segmentation on MRI with missing modalities. In: International Conference on Information Processing in Medical Imaging (2019)
24. Song, Z., Qi, Z., Wang, X., Zhao, X., Shen, Z., et al.: Uni-COAL: A unified framework for cross-modality synthesis and super-resolution of MR images. Expert Systems with Applications **270**, 126241 (2025)
25. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
26. Wang, X., Cui, Y., Wang, J., Zhang, F., Wang, Y., Zhang, X., Luo, Z., Sun, Q., Li, Z., Wang, Y., et al.: Multimodal learning with next-token prediction for large multimodal models. Nature pp. 1–7 (2026)

27. Wu, R., Wang, H., Chen, H.T., Carneiro, G.: Deep multimodal learning with missing modality: A survey. Transactions on Machine Learning Research (2026)
28. Xie, J., Darrell, T., Zettlemoyer, L., Wang, X.: Reconstruction alignment improves unified multimodal models. In: International Conference on Learning Representations (2026)
29. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
30. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review (2024)
31. Zhang, X., Ou, N., Doga Basaran, B., Visentin, M., Qiao, M., et al.: A foundation model for lesion segmentation on brain MRI with mixture of modality experts. IEEE Transactions on Medical Imaging **44**(6) (2025)
32. Zhao, F., Zhang, C., Geng, B.: Deep multimodal data fusion. ACM Computing Surveys **56**(9), 1–36 (2024)
33. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: RaTEScore: A metric for radiology report generation. In: Conference on Empirical Methods in Natural Language Processing (2024)