Conditional Latent Diffusion Models for Irregularly Spaced Longitudinal Radiological Data

Nabil Mouadden[1], Othmane Laousy[1], Rafael Marini[2], Valentin Ong[2], Marie-Pierre Revel[2], Guillaume Chassagnon[2], Stergios Christodoulidis[1], Maria Vakalopoulou[1,3]

[1]CentraleSupelec, IHU-National PRecISion Medicine Center in Oncology
[2] Department of Radiology, Hôpital Cochin, AP-HP. Centre Université Paris Cité
[3] Archimedes/Athena RC

Abstract. Longitudinal medical image studies often involves multiple scans of the same patient taken at different times, potentially with different modalities such as (2D vs. 3D volumetric medical imaging). In this work, we propose a single diffusion-based framework that can predict future embeddings of imaging data for predefined time points. Our approach uses a *universal vision encoder*, able to ingest either 2D or 3D scans, combined with a temporal transformer to fuse embeddings across multiple timepoints. A conditional *latent diffusion model* then produces the future output in latent space encoding the longitudinal information of the patient. We challenged our method in two crucial tasks involving radiological imaging: *(1)* predicting future pathology in the form of segmentation masks, exemplified by Interstitial Lung Disease (ILD) progression on 3D chest CT scans of Systemic Sclerosis (SSc) patients, and *(2)* generating radiology reports that incorporate prior imaging context, exemplified by longitudinal chest X-rays from MIMIC-CXR. Results indicate that this unified diffusion approach outperforms existing baselines in both pixel-level forecasting and report generation, highlighting its versatility and effectiveness for longitudinal medical imaging.

Keywords: Longitudinal Imaging · Diffusion Models · 2D/3D Vision · ILD · Radiology Reports

1 Introduction

Longitudinal medical image studies involve comparing multiple imaging data of the same patient captured at different time points that are not necessarily regularly spaced. Such studies are essential for assessing disease progression and thus making informed treatment decisions. Despite their importance, especially for managing chronic diseases, the development of longitudinal computational tools aiming to assist the clinical routine is challenged by the severe data scarcity, high variability and large dimensionality of medical images. In this study, we propose a novel deep learning method based on a conditional latent diffusion

model (CLDM) that generates future latent representations at predefined time points using multiple irregularly spaced past latent representations. These generated representations, when combined with task-specific decoders, can be used for a variety of applications, including *(1)* forecasting future disease state and *(2)* generating radiology reports that integrate prior and current findings. We evaluated the performance of the proposed method on a public and a real-world in-house clinical dataset, demonstrating its ability to outperform existing baseline models. Our findings suggest that the proposed pipeline holds promise as a valuable tool for clinicians in identifying possible future states and, ideally, improving their treatment options.

The contribution of this work is threefold; *(i)* we propose a unified conditional latent diffusion model that leverages a single architecture for processing 2D and 3D radiological data, *(ii)* in contrast to the current literature, our method is agnostic to the number of scans available per patient, and their temporal distribution which proves to be useful for clinical practice when the number of follow-ups varies for each patient, *(iii)* we introduce a temporal transformer based on the self-attention mechanism to model the temporal dynamics allowing the diffusion model to weigh the importance of different parts of the input sequence differently. Through experiments in two different datasets, we show that the proposed diffusion framework can tackle diverse tasks and data dimensionalities found in longitudinal imaging scenarios.

2 Related Work

Universal 2D/3D Vision Encoders. Universal pretrained encoders that can handle both 2D and 3D data with minimal architectural changes [17,4] have been introduced recently. Such unified models have demonstrated strong performance in a range of tasks, from planar chest X-ray classification to volumetric CT segmentation, by benefiting from a more generalizable representation space. Our pipeline leverages this idea of a universal encoder, allowing the same backbone to ingest 2D or 3D radiological data.

Diffusion Models in Medical Imaging. Diffusion models have recently shown promising results in various medical imaging tasks, including image synthesis, reconstruction, and implicit ensemble segmentation [7,11,16,2]. In practice, they learn a forward process that gradually corrupts data with noise and a reverse process that denoises the data in iterative steps, generating high-quality samples or reconstructions [12]. This framework has proven useful for handling complex, high-dimensional medical images, as it can capture subtle anatomical variations in a principled, probabilistic manner. Nevertheless, directly integrating temporal information into diffusion approaches remains underexplored, motivating our conditional latent diffusion design for multi-time scenarios.

Multi-Time Fusion in Longitudinal Analysis. While many deep learning models for medical imaging focus on single scans, longitudinal datasets offer richer context by capturing how a patient's condition evolves [8,10]. Conventional approaches sometimes concatenate features across time or adopt recur-

rent neural networks to exploit temporal cues. More recent methods employ transformer architectures to handle irregular intervals or missing scans, providing improved robustness in progression analysis. However, these works typically focus on either segmentation or classification, tasks, rarely extending to generative forecasting. Our approach differs by proposing a generative forecasting of the future embedding space, able to adapt to a variety of problems.

Automated Radiology Report Generation. Automating radiology reporting has long been pursued to reduce radiologist workload and ensure consistent quality of reports [14,13]. Early systems employed recurrent architectures [5]. More recently, transformer-based methods, such as R2Gen [1], enhanced coherence and factual correctness by leveraging attention over image features. More advanced frameworks incorporate prior reports or memory modules to highlight changes [20,3]. Recent diffusion-based text generators (e.g., [15]) refine textual embeddings in parallel, reducing errors that can accumulate in autoregressive decoders. By integrating a temporal transformer and a diffusion text decoder, we address the challenge of describing multi-exam findings more accurately than single-exam or purely autoregressive approaches.

Overall, while prior works have separately explored diffusion-based generation, multi-time fusion, and universal encoders for either 2D or 3D tasks, our framework merges these into a single, novel, efficient and modular pipeline.

3 Methods

Our approach is built around three key components: a universal vision encoder capable of ingesting either 2D or 3D scans, a temporal transformer that fuses multiple prior exams into a single representation, and a conditional latent diffusion module that produces the future latent embedding. Depending on the final decoder, the system outputs either a segmentation mask or a textual report. An overview of the entire method is presented in Fig. 1.

Universal Vision Encoder for 2D/3D Data. We use UniMiSS [18], a pretrained transformer based vision backbone, which processes medical images in either 2D or 3D form. We use this powerful model to extract embeddings $z_{d_i} \in \mathbb{R}^d$ for the image acquired on date d_i. This design enhances modularity since both 2D and 3D modalities are handled by a unified architecture, eliminating the need for separate specialized encoders.

Temporal Transformer for Longitudinal Fusion. After obtaining embeddings $\{z_{d_0}, z_{d_1}, \ldots, z_{d_k}\}$ from multiple prior exams, we assign each exam date d_i a trainable time embedding $d_i^t \in \mathbb{R}^d$. This vector is initialized randomly and updated via backpropagation, allowing the model to learn how to position each exam in the temporal axis. We then concatenate z_{d_i} with d_i^t into a single token u_{d_i}, and feed the token sequence $\{u_{d_0}, \ldots, u_{d_k}\}$ into a temporal transformer that uses self-attention to capture inter-exam relationships. We also insert a special `[CLS]` token, whose final hidden state $s_n \in \mathbb{R}^d$ summarizes the entire sequence, reflecting how the patient's condition evolved over $\{d_0, \ldots, d_k\}$. This design inherently accommodates variable scan counts and irregular intervals, as

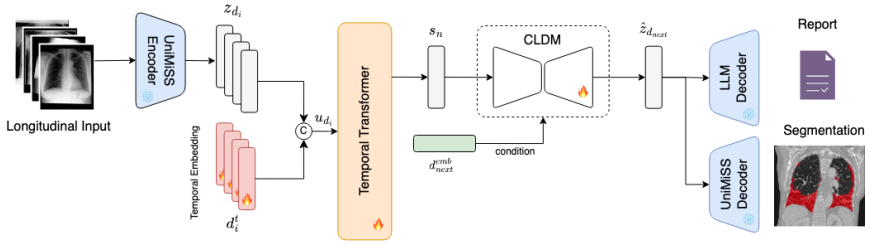

Fig. 1: Overview of our pipeline. 2D/3D CT scans are taken as input and embedded, these embeddings together with the date embeddings are combined into an full sequence representation and provided to the conditional diffusion model for future latent representation prediction. Task specific decoders can be then used to retrieve the final output.

the transformer attends to each time embedding based on its learned representation of proximity or distance in the patient timeline.

Conditional Latent Diffusion Model. Given the temporal context s_n and a target date embedding $d_{\text{next}}^{\text{emb}}$, we apply a latent diffusion model [12] to predict the future latent $\hat{z}_{d_{\text{next}}}$. During training, we have access to the ground-truth "future" embedding $z_{d_{\text{next}}}$, which we obtain by encoding the actual future data. We progressively add noise to $z_{d_{\text{next}}}$ over T steps in a forward diffusion. A U-Net style denoising network then learns to reverse it. At each step t, the network is conditioned on s_n, $d_{\text{next}}^{\text{emb}}$, and t, and predicts the noise to remove from z_t.

Training Procedure and Losses. We train our pipeline by jointly optimizing a *diffusion denoising objective* and a *task-specific reconstruction loss* depending on the decoder. Let θ denote the parameters of our diffusion model and temporal transformer. The overall loss that is minimized is summarized as,

$$\mathcal{L}(\theta) \;=\; \mathcal{L}_{\text{diff}}(\theta) \;+\; \lambda\,\mathcal{L}_{\text{task}}(\theta), \tag{1}$$

where $\lambda > 0$ is a hyperparameter that balances accurate denoising against the task specific loss depending on the decoder. We performed a small grid search over $\lambda \in \{0.1, 0.5, 1.0, 2.0\}$ on a validation subset and found that $\lambda = 1.0$ typically yielded the best balance between diffusion accuracy and final output performance. Consequently, we fix this value for all experiments reported in this paper. Using this design, we simultaneously teach the diffusion U-Net to *reconstruct* future embeddings and to produce final outputs that match ground-truth masks or reports. We use the default forward and reverse denoising process of the diffusion model and mean square loss for the $\mathcal{L}_{\text{diff}}(\theta)$. For the $\mathcal{L}_{\text{task}}(\theta)$, we used classical cross-entropy loss for the pixel-wise semantic segmentation as well as the report generation. To make the training more robust to random time intervals, during each training epoch, we randomly pick a future time d_{next} for each patient, ensuring a wide distribution of intervals. This induces robustness to variable follow-up durations, so the model generalizes better to time gaps not explicitly seen in the training set.

Inference and Few-Shot Generalization. At test time, we do not have ground-truth data for the future date. Instead, we sample an initial noise vector

in latent space, then run the learned denoising model in reverse for T steps, conditioned on the summary vector s_n and $d_{\text{next}}^{\text{emb}}$. The result is a predicted future latent $\hat{z}_{d_{\text{next}}}$ which is then given to a pretrained decoder. For this study, we tested one decoder for semantic segmentation of ILD disease and one for text generation. This procedure accommodates varying amounts of historical data: if the patient has many prior scans, the temporal transformer integrates all of them; if only one or two are available, the model still provides a forecast. The system naturally generalizes to unseen intervals or fewer time points, providing a few-shot or zero-shot capacity in real clinical scenarios.

Implementation Details. We used UniMiSS [18] encoder, setting the latent dimension to $d = 768$. A temporal transformer with four layers of multi-head attention and feed-forward blocks processes the multi-time context, producing s_n. Our diffusion stage uses a U-Net [12] enhanced by cross-attention on $(s_n, d_{\text{next}}^{\text{emb}})$ and sinusoidal timestep embeddings; we typically choose $T = 1000$ steps for training, with the option to reduce it at inference via accelerated sampling. For segmentation, we employ the lightweight UniMiSS decoder that mirrors the encoder's resolution hierarchy, using transposed convolutions to reconstruct the mask from $\hat{z}_{d_{\text{next}}}$. We fine-tuned this decoder for ILD segmentation. For text generation, we use BioGPT [9] as a generative language model that cross-attends to $\hat{z}_{d_{\text{next}}}$, decoding radiology reports autoregressively. All modules are trained end-to-end with separate final decoders for each domain. We use AdamW optimization and tune hyperparameters (e.g. diffusion weights, cross-entropy weights) via validation. During inference, only the relevant task decoder (segmentation or text) is needed, whereas the encoder, temporal transformer, and diffusion U-Net remain shared across both tasks.

4 Results and Discussion

4.1 ILD Progression with Volume-Based Evaluation

We first apply our approach to forecasting *future* ILD progression in patients with Systemic Sclerosis (SSc). For this task, we used a private dataset of 1,359

Table 1: Evaluation of fibrotic *volume progression* in SSc-ILD, using Dice (higher is better) and adjusted Volume Difference (αVD; lower is better).

Method	**Dice**	**αVD**
Naive (Copy Last Volume)	0.65	0.52
No Diffusion (Deterministic)	0.27	1.89
SADM [19] (all Scans)	0.34	1.40
Ours (without diffusion, all Scans)	0.27	1.89
Ours (1 Scan)	0.55	0.93
Ours (3 Scans)	0.62	0.53
Ours (All Scans)	**0.76**	**0.30**

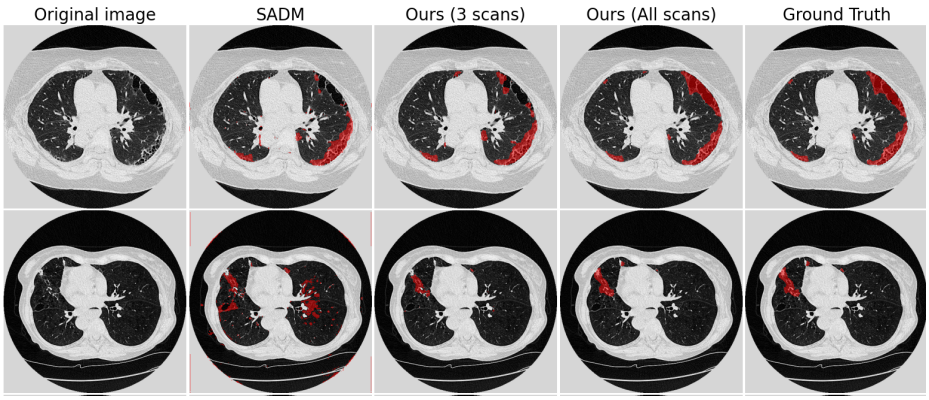

Fig. 2: Qualitative comparison between different configurations of our approach and SADM. Two 2D slices in axial view from different patients are presented together with the disease regions superimposed in red color.

chest CT scans, all annotated for fibrotic tissue, collected over 16 years from 230 patients. Of these patients, we kept 30 as an independent test set. Unlike single-time segmentation, our goal is to predict how fibrotic regions evolve at a forthcoming date.

Evaluation Metric. Since stable fibrotic tissue can inflate overlap metrics (such as Dice coefficient), we emphasize a *volume-based* measure to capture genuine progression. Specifically, we adopt the *adjusted Volume Difference*: aVD = $\frac{|V_{\text{pred}} - V_{\text{gt}}|}{V_{\text{lung}}}$, where V_{pred} and V_{gt} are predicted vs. ground-truth fibrotic volumes, and V_{lung} is total lung volume. Lower aVD implies better alignment in forecasting the fibrotic burden and this is a metric added to the classical dice metric.

Baselines. We compare our method against the state of the art Sequence-Aware Diffusion Model (SADM) [19] for generating longitudinal medical images and two baselines including the *naive* which replicates the last known fibrotic volume, ignoring potential new growth and the *no diffusion* which is a simpler deterministic version of our method that uses a direct MLP regressor from the temporal transformer to the future embedding. This tests whether generative noise-refinement significantly helps.

Results. Table 1 provides the aVD and Dice indexes for our different methods, benchmarking as well the importance of the use of different scans. Our proposed method, using all the available scans yields a dice of 0.76 and an aVD of 0.30, outperforming both the naive baseline and the deterministic approach, while it shows superiority with respect to the previous state of the art SADM model. Additionally, Figure 2 presents two axial slices from different patients, illustrating how our method performs when using only *3 scans* versus *all available scans*, compared to the ground truth. The first column shows the original CT slice, with the fibrotic region in subsequent columns highlighted in red.

Table 2: Radiology report generation on MIMIC-CXR (higher is better). "+Prior" indicates usage of a previous exam.

Method	+Prior	BLEU-4	ROUGE-L	CheXpert F1
R2Gen [1]	No	0.14	0.28	0.27
ControlDiff	No	0.13	0.28	0.28
Ours (Diffusion)	No	**0.15**	**0.30**	**0.31**
Longitudinal Transf. [20]	Yes	0.16	0.33	0.35
Ours (Diffusion)	Yes	**0.17**	**0.34**	**0.38**

4.2 Longitudinal Radiology Reporting (2D X-ray)

We apply the same pipeline to the publicly available MIMIC-CXR[6], which couples longitudinal chest X-rays with radiology reports. Our method encodes prior and current images with the universal encoder, merges them via a temporal transformer, then uses a diffusion-based text decoder to produce the new report.

Evaluation Metric. We measure BLEU-4 and ROUGE-L to capture n-gram overlap with reference text and overall summary-like matching, respectively. Although these metrics gauge fluency and surface similarity, they do not guarantee clinical correctness. To address this limitation, we also use a CheXpert label F1 score, whereby an automated labeler extracts findings (e.g., "effusion," "cardiomegaly") from both the generated and the ground-truth reports. This produces a label-based precision, recall, and F1 measure, reflecting how accurately the model reports key pathological states. Such *clinically oriented* metrics better reflect the actual diagnostic utility of the reports, albeit they can appear lower than purely linguistic scores due to the strictness of label matching.

Baselines. We compare to three different approaches including the R2Gen [1], the ControlDiff, a diffusion text generator *without* explicit multi-time conditioning, clarifying the benefit of temporal embeddings and the Longitudinal Transformer [20] which incorporates prior scans, but decodes autoregressively.

Results. Table 2 compares BLEU-4, ROUGE-L, and CheXpert F1, with or without using a prior exam for the different methods. Our approach exceeds all baselines in text quality and clinical correctness, especially when prior scans are available. Moreover, Table 3 includes two examples of reports generated by our method and the longitudinal AR baseline. Once more, our proposed model using the prior scan generates more complete and accurate reports, encoding the disease evolution.

5 Discussion and Conclusion

In this paper, we presented a single diffusion-based pipeline for two key longitudinal imaging tasks: forecasting future pathology in 3D CT scans (as demonstrated with SSc-ILD) and generating radiology reports over repeated 2D chest X-rays. Our approach exploits a universal vision encoder to handle 2D/3D inputs, a

Table 3: Examples of generated reports (paraphrased for brevity) across different methods. **GT** = Ground Truth. "with prior" denotes models that incorporate the previous exam.

Example 1:
Ground Truth (GT): *"Large left pleural effusion has significantly decreased since prior exam after thoracentesis. Moderate bibasilar atelectasis. Heart size unchanged."*
Ours (with prior): *"Marked decrease in left pleural effusion compared to prior exam. Remaining small left effusion persists. Bibasilar opacities likely atelectasis. Heart size is stable."*
Ours (no prior): *"Large left pleural effusion is present. Bibasilar opacity consistent with atelectasis. Heart size is mildly enlarged."* (*Fails to mention improvement, focusing on current findings only.*)
Longitudinal AR Baseline (with prior): *"Left effusion is improved from prior exam. Left pleural effusion seen. Basilar atelectasis present. Cardiomegaly unchanged."* (*Acknowledges improvement but redundantly repeats effusion.*)
Example 2:
Ground Truth (GT): *"Mild cardiomegaly with unchanged patchy pulmonary edema. Small bilateral effusions, no significant interval change."*
Ours (with prior): *"Mild cardiomegaly and pulmonary edema with no significant change compared to prior. Small bilateral pleural effusions remain stable."*
Ours (no prior): *"Mild cardiomegaly with patchy pulmonary edema. Small bilateral pleural effusions are present."* (*Describes findings but omits explicit comparison due to lacking prior context.*)
Longitudinal AR Baseline (with prior): *"Heart size and pulmonary edema similar to prior. Bilateral effusions unchanged. Mild cardiomegaly."* (*Conveys similar info but slightly less fluent ordering.*)

temporal transformer to fuse multi-time embeddings, and a conditional diffusion process that iteratively refines a "future" latent representation, decoded as either a segmentation mask or text. Empirical results show notable improvements in volume-based ILD progression metrics and language/clinical metrics for multi-exam reporting. While partial stability of disease regions can complicate overlap-based metrics, our volume-based measure underscores the benefit of generative diffusion in capturing progression. Similarly, the text generation experiments validate that parallel refinement can yield more coherent references to changes than standard autoregressive decoders. Future directions may include difference-based annotations for new fibrotic tissue, synergy between predicted masks and textual descriptions, and integration of external clinical information for even richer forecasting. We believe that unifying multi-dimensional imaging with a single generative pipeline opens the door for more holistic, efficient solutions to the challenges of real-world longitudinal imaging.

Acknowledgments. This work has been partially supported by ANR-23-IAHU-0002, ANR-21-CE45-0007, ANR-23-CE45-0029, and by the project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGen- erationEU Program. This work was performed using HPC resources from the *Mesocentre* computing center of CentraleSupelec.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, T., et al.: Generating radiology reports via memory-driven transformer. MIDL (2020)
2. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated MRI. Medical Image Analysis **80**, 102479 (2022)
3. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
4. He, X., Yang, Y., Jiang, X., Luo, X., Hu, H., Zhao, S., Li, D., Yang, Y., Qiu, L.: Unified medical image pre-training in language-guided common semantic space. arXiv preprint arXiv:2311.14851 (2024)
5. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia (Jul 2018). `https://doi.org/10.18653/v1/P18-1240`, `https://aclanthology.org/P18-1240/`
6. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., et al.: MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
7. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis p. 102846 (2023)
8. Li, H., Zhang, H., Johnson, H., Long, J.D., Paulsen, J.S., Oguz, I.: Longitudinal subcortical segmentation with deep learning. Medical Imaging 2021: Image Processing **11596**, 115960D (2021)
9. Luo, R., Sun, L., Xia, Y., Qin, B., Liu, T.: Biogpt: Generative pre-trained transformer for biomedical text generation and mining. arXiv preprint arXiv:2210.10341 (2022)
10. Mayfield, J.D., Murtagh, R., Ciotti, J., Robertson, D., El Naqa, I.: Time-dependent deep learning prediction of multiple sclerosis disability. Journal of Imaging Informatics in Medicine **37**(6), 3231–3249 (2024)
11. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
13. Sirshar, M., Paracha, M.F.K., Akram, M.U., Alghamdi, N.S., Zaidi, S.Z.Y., Fatima, T.: Attention based automated radiology report generation using cnn and lstm. PLoS ONE **17**(1), e0262209 (2022)

14. Sloan, P., Clatworthy, P., Simpson, E., Mirmehdi, M.: Automated radiology report generation: A review of recent advances. IEEE Reviews in Biomedical Engineering (2024)
15. Tang, F., et al.: Abnormal semantic diffusion for radiology report generation. AAAI (2024)
16. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: Medical Imaging with Deep Learning (MIDL). pp. 1336–1348 (2022)
17. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unified 2d and 3d pre-training for medical image classification and segmentation. arXiv preprint arXiv:2112.09356 (2021)
18. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)
19. Yoon, J.S., Zhang, C., Suk, H.I., Guo, J., Li, X.: Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In: Frangi, A., de Bruijne, M., Wassermann, D., Navab, N. (eds.) Information Processing in Medical Imaging. pp. 388–400. Springer Nature Switzerland, Cham (2023)
20. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–198. Springer (2023)