

# CEST MRI Z-Spectra Reconstruction Using Transformer-based Deep Learning Models and Under Sampled Spectrum Signatures

Hemanth Pasupuleti <sup>a,b</sup>, Kelden Pruitt <sup>a,b</sup>, Elena Vinogradov <sup>c</sup>, Fang Frank Yu <sup>c</sup>, Baowei Fei <sup>a,b,c</sup>

<sup>a</sup> Center for Imaging and Surgical Innovation, University of Texas at Dallas, Richardson, TX

<sup>b</sup> Department of Bioengineering, University of Texas at Dallas, Richardson, TX

<sup>c</sup> Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX

\* Corresponding author: bfei@utdallas.edu, Website: <https://fei-lab.org>

## ABSTRACT

Chemical exchange saturation transfer (CEST) contrast MRI provides unique molecular contrast by probing proton exchange between solutes and bulk water, but its clinical translation is limited by long acquisition times required to acquire finely sampled Z-spectra bringing a tradeoff between acquisition time and frequency offset responses. In this preliminary study, we investigated the performance of transformer-based deep learning models to reconstruct fully sampled CEST spectra. We first validated the approach using a lightweight model on simulated Z-spectra, followed by a scaled high-capacity transformer for human MRI data to handle the increased data scaling and spectral variability. We evaluate the model reconstruction quality on simulated Z-spectra and the human MRI data using root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and cosine similarity for regression analysis. We use structural similarity index measure (SSIM) and peak signal to noise ratio (PSNR) to evaluate image-level reconstruction quality. The transformer approach accurately recovered missing spectral information with an MAPE of 3.41% and RMSE of 0.0169. The reconstruction maintained high structural fidelity with a mean SSIM of 0.9658 and spectral fidelity with Cosine Similarity of 0.9951. These results validate that the scaled transformer architecture ensures robust generalization and accurately captures complex spectral dependencies in clinical data. These findings highlight the potential of deep learning-based reconstruction to accelerate CEST MRI acquisition supporting more efficient clinical protocols.

**Keywords:** Magnetic resonance imaging (MRI), chemical exchange saturation transfer (CEST), deep learning, transformer model, Z-spectra reconstruction

## 1. INTRODUCTION

Magnetic resonance imaging (MRI) plays a vital role in clinical diagnosis, offering high-resolution physiological and anatomical insights. Chemical exchange saturation transfer (CEST) is an advanced MRI contrast mechanism that allows detection of low-concentration molecules by leveraging the exchange of protons between exogenous or endogenous compounds and water [1]. A saturation RF pulse is applied at the frequency of the exchanging protons and water, and the resultant decrease in water intensity is observed. Often, multiple images are acquired as a function of the RF pulse frequency offset. The resulting Z-spectrum provides rich contrast of the tissue, making CEST highly valuable for tissue characterization, e.g. tumors monitoring [2]. Despite its potential, widespread adoption of CEST is hindered by its inherently slow acquisition process. The spectral information acquired depends on the sampling of the frequency offsets. Smaller step sizes provide more accurate spectral resolution but dramatically increase scan time, making clinical translation challenging. Conversely, larger step sizes reduce acquisition time but at the cost of losing critical spectral details, limiting diagnostic reliability.

In recent advances of deep learning, transformer model-based architectures have shown powerful capabilities in modelling sequential and spatial data [3,4]. These models are well-suited to reconstruct these Z-spectra points due to their ability to capture global and local relationship dependencies because of the attention mechanism. Hence, if we can make them learn the Z-spectra structure and potentially enabling accurate recovery of Z-spectra from under sampled acquisitions, deep learning may allow the use of larger step sizes without sacrificing experimental quality, thus accelerating CEST imaging for clinical use.

In this preliminary study, we investigated the application of one-dimensional transformer based deep learning models for reconstructing under sampled CEST Z-spectra. We evaluate this method on simulated CEST data and the real MRI data belonging to human patients on both regression and image-focused metrics to capture the individual Z-spectra reconstruction and the structural reconstruction information. Our objective is to demonstrate that deep learning can bridge the trade-off between acquisition speed and spectral sampling frequency, paving the way for more efficient deep learning-powered CEST MRI acquisition.

## 2. METHODS

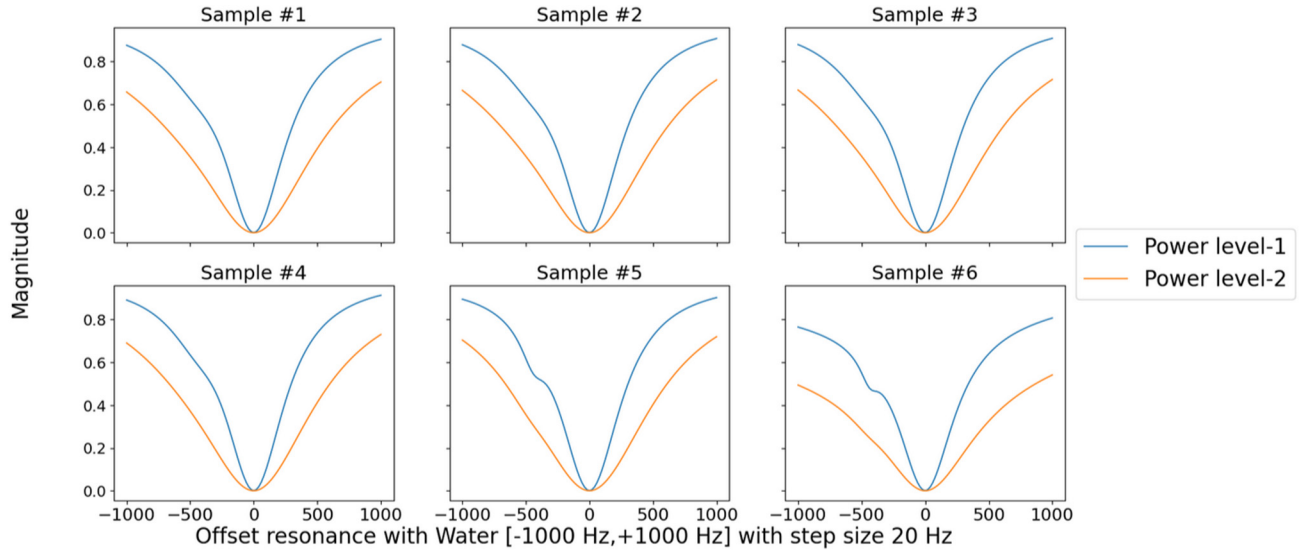
### 2.1 MRI spectrum simulation and data acquisition

#### CEST Simulated Data

For our initial work, we generated CEST simulated data using in-house developed MATLAB codes generating 1D Z-spectra CEST experiments using Bloch-McConnell (BM) equations. The simulated dataset contains 1D Z-spectra for a 4-pool model system at 3T. The 4 pools considered are: (a) water (observed in the experiment) at 0 parts per million (ppm), (b) amide group (-NH) at 3.5 ppm, (c) lipids giving so-called NOE (rNOE) at -3.4 ppm, and (d) MT pool (broad amorphous) at -2.4 ppm. At 3T, these 4 pools become: (a) water (0 Hz), (b) amide group (447.09 Hz), (c) rNOE (-434.316 Hz), and (d) MT (-306.576 Hz). Each simulation generates a Z-spectra for different: Concentrations, Exchange rates, and  $T2$  parameters for 3 pools excluding water. The output is water Z-magnetization which is proportional to the observed signal as a function of different parameters corresponding to physical properties of pools and exchange rate between them.

After running the simulations, the following properties are used in our final data structure: (1)  $T2_b$ ,  $T2_c$ , and  $T2_d$  are  $T2$  values of pools (b), (c), and (d), (2)  $Tau_b$ ,  $Tau_c$ , and  $Tau_d$  are inverse of exchange rate between (b), (c), (d) and (a), (3)  $M_b$ ,  $M_c$ ,  $M_d$  are different concentrations of pools corresponding to ratio of pools (b), (c), (d) to water (a), (4) frequency offset resonance with water, and finally (5) scales of  $B1$  used in experiment. The final data structure follows the format:  $(N_{T2_d}, N_{T2_c}, N_{T2_b}, N_{Tau_d}, N_{Tau_c}, N_{Tau_b}, N_{M_d}, N_{M_c}, N_{M_b}, N_{offset}, N_{scale})$  where N represents the number of values the parameter has, which becomes: (3, 3, 3, 3, 3, 3, 3, 3, 3, 101, 2). We have 3 samples for each parameter  $T2$ ,  $Tau$ , and  $M$ , while the  $N_{offset}$  is 101 because the Z-spectrum frequency ranges [-1000 Hz, +1000 Hz] with a step size of 20 Hz giving us a total 101 steps, this dimension contains the Z-spectra values corresponding to the offset step.

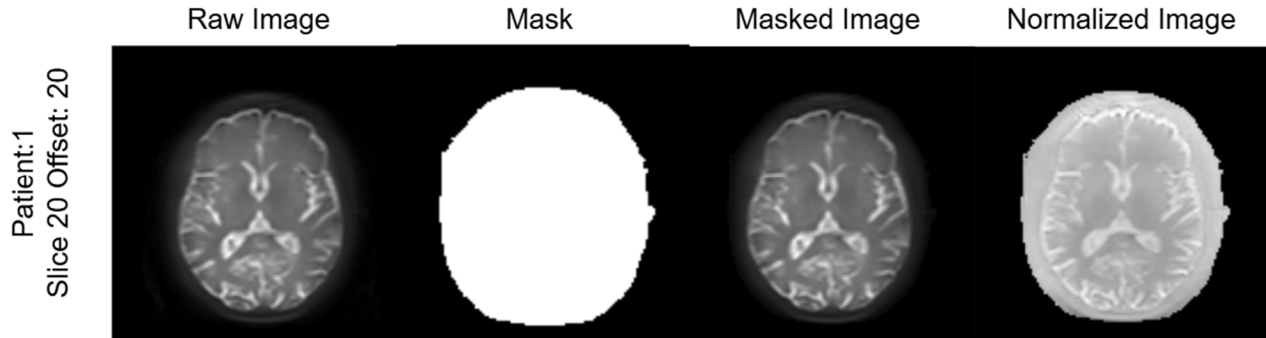
For machine learning training, we flatten this data and consider the Z-spectra with 101 steps and 2 power levels as the signal to be reconstructed with a sample size of 19,683. Our total dataset has dimensions of (19683, 101, 2), which we split into 60:20:20 training/validation/testing with training dataset of size (11809, 101, 2), validation and testing datasets of size (3937, 101, 2).



**Figure 1.** Sample Z-spectra from 1D CEST simulated data with 2 power levels. X-axis represents the offset resonance with water, and Y-axis represents the magnitude.

### Patient MRI Data

To evaluate the training paradigm on real world dataset, MR images of six normal human subjects were used in this study. The MRI DICOM images were processed to have dimensions of (128, 128, 46, 24) where the 24<sup>th</sup> column in the 4<sup>th</sup> dimension represents the normalization parameter which we can divide to normalize the spectra to [0, 1]. We preprocess these MRI images to remove background noise through thresholding across the normalization parameter, giving us a mask. We then use this mask to segment out the region of interest from the MRI data which is then normalized. The final processed MR image after normalization would be of dimension (128, 128, 46, 23). For training and testing we use leave one out cross validation (LOOCV) to train on the data of five patients and cross-validate on one patient data. We apply masking to get all the 1D spectra (the last dimension) from the MR images giving us an average of 356,612 samples per patient.



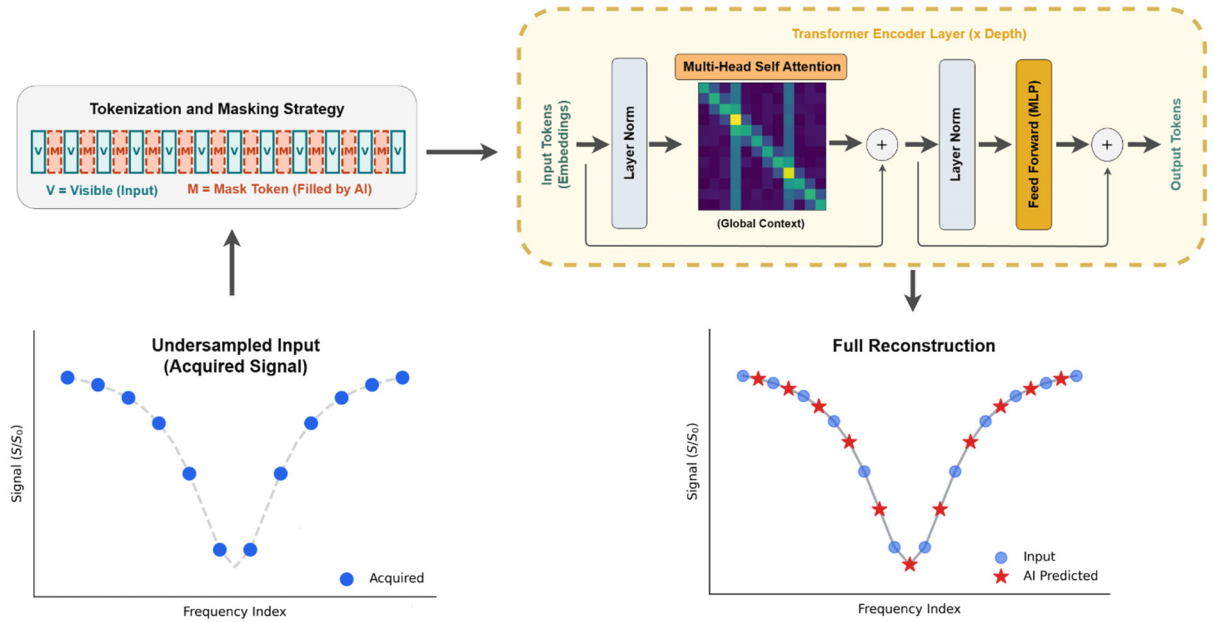
**Figure 2.** Sample patient MRI slice of one offset. The MRI is of dimension (128, 128, 46, 23), the image belongs to Slice 20 out of 46 giving us (128, 128, 23) then we take the 20th offset from channel to visualize the steps involved. Raw image is thresholded on normalization parameter to obtain the mask which is then used to remove background and then the final output is normalized to obtain the processed MR image.

### 2.2 1D Transformer Model

Transformer architecture based deep learning models have become popular due to their attention mechanism, which enables modelling of complex dependencies within input data [3,4]. While they are commonly applied to high-dimensional sequences such as text or images, their ability to learn global relationships is equally beneficial for lower-dimensional Z-spectra data due to their ability to generalize well across the dataset. Hence, we employed a one-dimensional (1D) transformer model to train and reconstruct the Z-spectra.

For the simulated CEST dataset, each Z-spectrum consists of 101 frequency offsets at two power levels, resulting in input dimensions of (101, 2). In contrast, the processed patient MRI data utilizes a specific acquisition protocol with 23 normalized frequency offsets at a single power level, resulting in input dimensions of (23, 1). These inputs were treated as sequential data and were mapped into higher-dimensional embedding space using linear projection layer while adding the positional embedding [3]. The embedded sequence is then processed by a series of transformer blocks composed of multi-head attention. Finally, a linear layer projects the high-dimensional space back into the original input dimensions to yield a reconstructed Z-spectrum.

The model was trained using a self-supervised masked prediction framework. During training, every other input element of the Z-spectrum was replaced with a learnable mask token, corresponding to the masking of alternate frequency offsets. This strategy simulates the doubling of the step size in the input sequence, requiring the model to predict the missing values from the available context. The masked sequence was then processed by the transformer and optimized to reconstruct the original Z-spectrum values at the masked positions. During inference, this same masking procedure simulates the acquisition of data with an increased step size and the model thus functions as a spectral super-resolution module, recovering dense Z-spectra from sparsely sampled inputs.



**Figure 3.** Illustration of the 1D Transformer-based reconstruction workflow. The process begins with an under-sampled input signal representing the acquired Z-spectrum. Mask tokens are inserted at positions corresponding to unacquired frequency offsets. This sequence is processed by the 1D Transformer encoder, which utilizes attention mechanisms to predict and fill in missing spectral details, resulting in a fully sampled Z-spectrum.

### 2.3 Model Training and Inference

We adopted a staged development strategy to address the distinct complexity requirements of simulated versus clinical data. All models were trained on a single NVIDIA RTX A6000 GPU (48 GB memory).

*Simulated Data Model Configuration:* For the initial validation on simulated CEST data, we utilized a lightweight transformer architecture. Given the controlled nature of simulation and available data samples, a compact model with approximately 450 trainable parameters was sufficient to validate the masked prediction framework. This model utilized a small MLP dimension of 8 and four encoder layers, optimized with a batch size of 64.

*Patient MRI Data Model Configuration:* To account for the large data volume and high spectral variability inherent in clinical scans, we significantly scaled the model architecture for the patient MRI data. Processing the flattened MRI volumes yields approximately 250,000 spectral samples from the masked region per patient. To capture the complex non-linear dependencies in this large-scale dataset, we implemented a high-capacity transformer with 9,464,321 trainable parameters. The scaled architecture featured an embedding dimension of 512, a depth of 6 transformer encoder layers, and 8 attention heads with a head dimension of 64. The MLP hidden dimension was expanded to 512. The input sequence length was fixed at 23, corresponding to the normalized frequency offsets.

The training followed the self-supervised masked prediction strategy described previously, where alternate input points are masked and the model was trained to reconstruct them. The model was optimized by minimizing the L1 loss, or Mean Absolute Error (MAE), defined as:

$$L1 = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

Where  $i$  ranges across all masked spectrum values within a single Z-spectra,  $y_i$  represents the ground truth spectrum values that were masked and  $\hat{y}_i$  represents the Z-spectrum values predicted by the model.

We trained both the simulated CEST and MRI models using AdamW optimizer [5] with a learning rate of  $1 \times 10^{-4}$ , reducing every 5 epochs upon plateau by a factor of 0.9. The high-capacity model is trained for 20 epochs, while the smaller simulated data model is trained for 200 epochs.

## 2.4 Evaluation Metrics

To assess the model performance, we employed both regression metrics (to evaluate the spectra reconstruction) and image quality metrics (for reconstructed MRI images). For regression metrics, we used Mean Absolute Error (MAE) (Eq.1), Mean Squared Error (MSE) (Eq.2), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) (Eq.3) and Cosine Similarity (Eq.4). MSE measures the average squared difference between the predicted and ground-truth values. RMSE is calculated as the square root of MSE. MAE measures the average absolute difference between the predicted and ground-truth. MAPE expresses the average prediction error as a percentage of true value. Cosine Similarity calculates the cosine angle between the two vectors, the closer they are, the better the prediction. For image-focused reconstruction evaluation, we consider splitting the MRI into images of dimensions (128, 128) for each spectra giving us total images of 1058 (46\*23) per MRI image. We then evaluated these images using Structural Similarity Index Measure [6] and PSNR (Eq.5) metrics. SSIM measures the structural properties and local patterns while PSNR calculates the logarithmic measure of ratio between maximum possible signal intensity and reconstruction error.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3)$$

$$\text{Cosine Similarity} = \frac{\mathbf{y} \cdot \hat{\mathbf{y}}}{\|\mathbf{y}\| \cdot \|\hat{\mathbf{y}}\|} \quad (4)$$

$$PSNR(\text{Image}) = 10 \cdot \log_{10} \left( \frac{MAX_{\text{Image}}^2}{MSE(\text{Image})} \right) \quad (5)$$

Where for all equations,  $i$  ranges across the spectrum values for a given Z-spectra, and for Eq.4 each Z-spectra is considered as a vector and in Eq.5,  $MSE(\text{Image})$  represents the average square of differences between the pixel intensities of real image and predicted image each with dimensions of (128, 128).

## 3. RESULTS

### 3.1 Evaluation of Simulated Data

Using the lightweight transformer configuration (~450 parameters), the model successfully validated the spectral super-resolution hypothesis. Table 1 summarizes the regression metrics for the model trained on the simulated Z-Spectra. The model achieved strong performance, with a mean of 0.0106 and MAE of 0.0681 with standard deviations (STD) of 0.026 and 0.075. MAPE has a mean of 13.08% and 21.84% STD. Cosine similarity is also very high at 0.9984 with very low standard deviation. However, upon further analysis, we observed the presence of small subset of outlier cases (~14%), primarily arising from the effect of simulated noise and signal variability. Hence, we also reported the metrics after removing these outliers from the dataset which led to drastic changes in metrics such as MAPE where it improved from 13% mean with 22% STD to 6.25% mean and 4.05% STD. We removed these outliers by calculating 25th percentile (First Quartile  $Q1$ ), median (Second Quartile  $Q2$ ), and 75th percentile (Third Quartile  $Q3$ ). Then, interquartile range (IQR) is simply the difference between  $Q3$  and  $Q1$ . Every sample that does not fall in range  $[Q1 - (1.5 \times IQR), Q3 + (1.5 \times IQR)]$  is considered as an outlier and is thus removed.

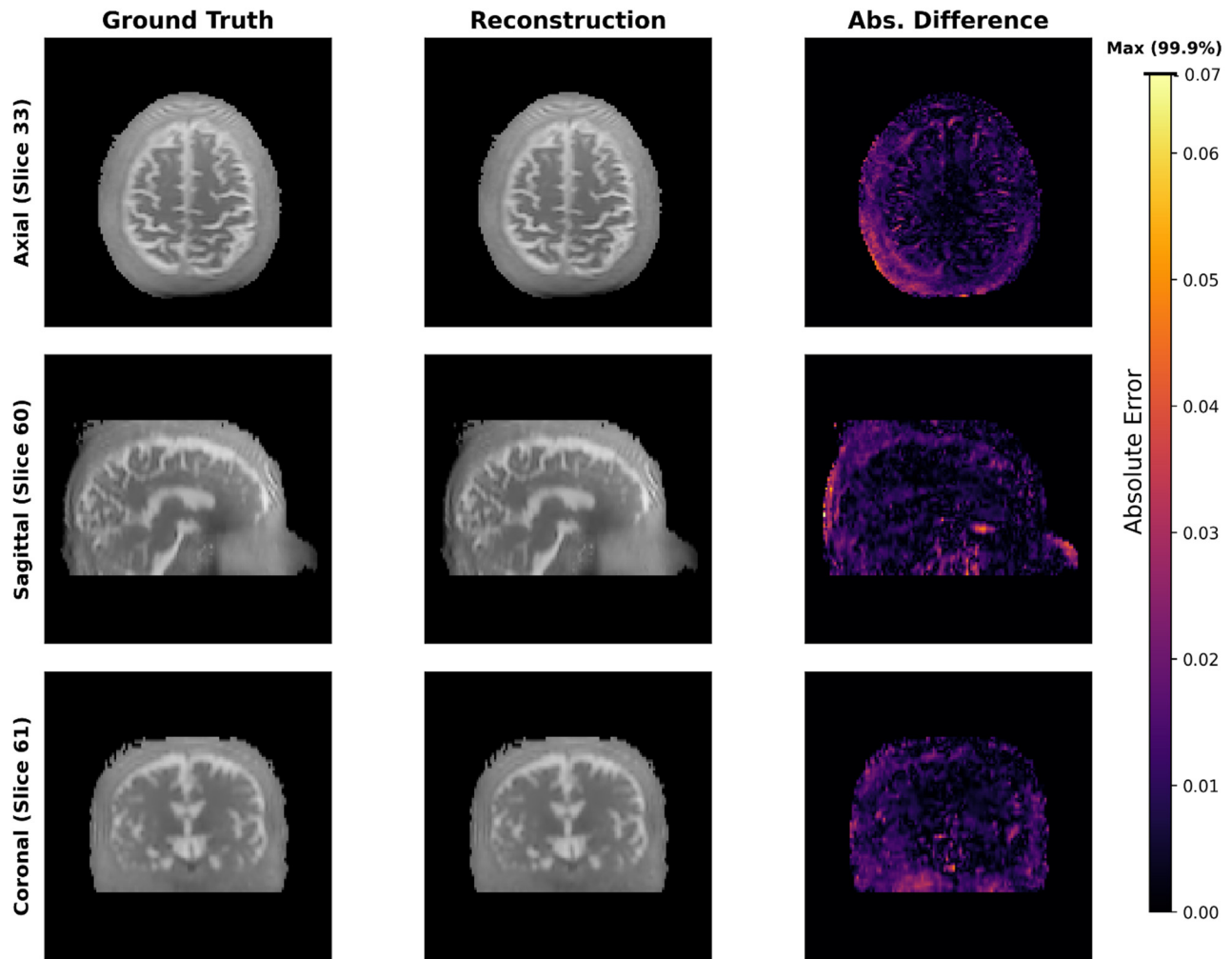
**Table 1.** Evaluation of MSE, MAE, MAPE, and Cosine metrics on testing data for simulated data.

	Mean	STD	Outlier %	Mean w/o outliers	STD w/o outliers
MSE	0.0106	0.0260	14.24	0.0027	0.0027
MAE	0.0681	0.0751	8.86	0.0475	0.0302
MAPE	13.08	21.84	13.84	6.25	4.05
Cosine	0.9984	0.0036	15.47	0.9996	0.0004

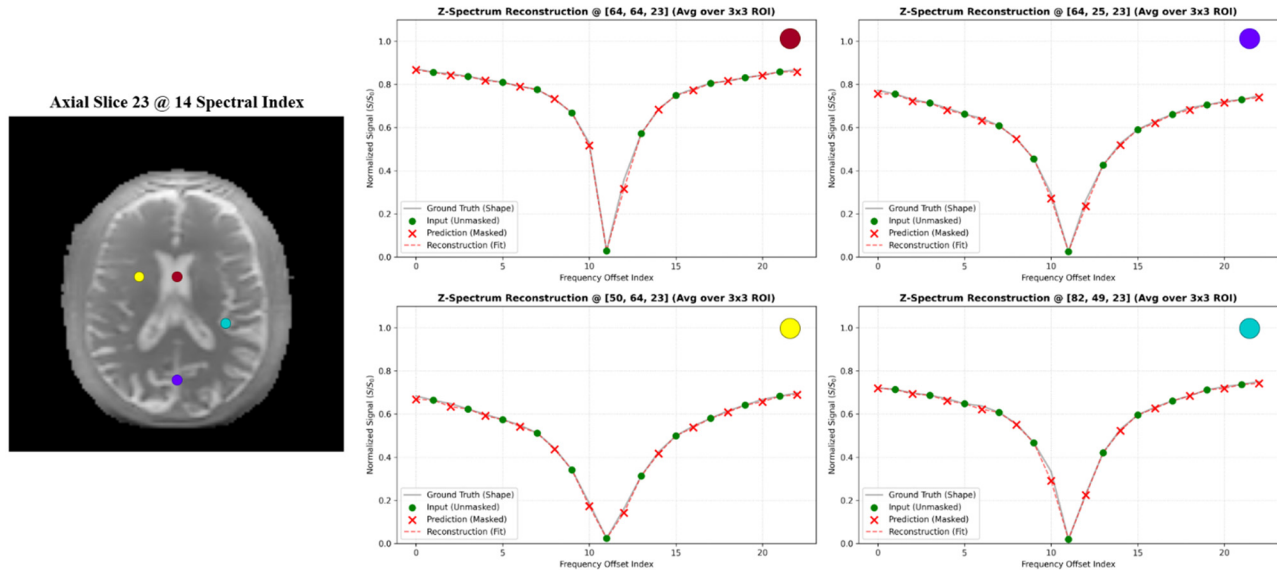
### 3.2 Evaluation of Patient Data

Figure 4 illustrates the spatial reconstruction quality for Patient 1 across three anatomical planes. The absolute difference maps (third column) reveal minimal residual errors, with the maximum error (99.9<sup>th</sup> percentile) capped at 0.07. This confirms that the model effectively eliminates aliasing artifacts typically associated with under sampling while preserving

high-frequency tissue details. Furthermore, Figure 5 presents spectral reconstruction at specific voxel locations. To ensure robust evaluation, signals were averaged over a 3x3 pixel window. The model accurately interpolates the missing frequency offsets (marked by red 'x'), closely adhering to the ground truth Z-spectrum curve (grey line) even in regions with steep signal dips, validating the transformer's ability to capture complex spectral dependencies.



**Figure 4.** Qualitative evaluation of spatial reconstruction on Patient 1. The columns display the Ground Truth, the Model Reconstruction, and the Absolute Difference map across Axial (Top), Sagittal (middle), and Coronal (bottom) views. The difference map highlights the error magnitude, with the color bar scaled to the maximum error (99.9th percentile) of 0.07 for the patient, demonstrating minimal structural deviation.



**Figure 5.** MRI Spectral Reconstruction qualitative evaluation. Representative Z-spectra reconstructions are shown for selected regions of interest (ROIs) within the Axial slice of patient 1. The plotted signal represents the average intensity over a 3x3 window to mitigate local noise. The red ‘x’ markers indicate the model’s prediction at the masked positions, while the dashed red line shows the full reconstruction fit against the ground truth (grey line).

We employed Leave-One-Out Cross-Validation (LOOCV) to evaluate the model’s performance on real-world data. In each fold, the model was trained on five patients and tested on the remaining patients. Table 2 details the performance metrics for each individual patient split. The model demonstrated consistent performance with Structural Similarity Index Measure (SSIM) scores ranging from 0.9502 to 0.9770 and Cosine Similarity consistently above 0.98.

**Table 2.** Patient-wise Cross-Validation Results. Evaluation metrics for each test fold (Patient 1-5) using the Leave-One-Out method.

Patient #	SSIM	PSNR (dB)	RMSE	MAPE (%)	Cosine
1	0.9770	34.79	0.0149	2.83	0.9970
2	0.9502	31.57	0.0151	2.90	0.9980
3	0.9668	32.33	0.0162	4.10	0.9847
4	0.9718	32.33	0.0225	4.25	0.9983
5	0.9634	34.17	0.0160	3.00	0.9977

To assess the overall generalizability of the framework, we aggregated all five cross-validation folds. As shown in Table 3, the proposed method achieved a mean SSIM of 0.9658 ( $\pm$  0.0091 STD) and a Peak-Signal-to-Noise-Ratio (PSNR) of 33.04 dB ( $\pm$  1.23 STD) across multiple patients, indicating strong preservation of spatial structure. Quantitatively, the spectral reconstruction error was low, with a mean RMSE of 0.0169 and a MAPE of 3.41%.

**Table 3.** Aggregated Cross-Validation Results. Mean and Standard Deviation (STD) calculated across all five patient folds.

Metric	Mean	STD
SSIM	0.9658	0.0091
PSNR (dB)	33.04	1.2258
RMSE	0.0169	0.0028
MAPE	3.4148	0.6234
Cosine	0.9951	0.0052

## 4. DISCUSSION AND CONCLUSION

In this study, we developed and evaluated transformer-based architecture utilizing a self-supervised masking mechanism for accelerating CEST-MRI spectral reconstruction. To achieve this, we initially created a synthetic CEST dataset simulating MRI spectrum signatures and strategically masked Z-spectra input to simulate an increased step size. By masking every other frequency offset to effectively double the step size, we utilized the transformer model as a spectral super-resolution module. We then evaluated the architecture's performance on clinical data consisting of MRIs from six patients. The results demonstrated that the model architecture learned to robustly interpolate missing points with a mean RMSE of 0.0169 and MAPE of 3.41% on real human subject data. Furthermore, the image-focused metrics SSIM and PSNR indicate that the model preserves the spatial structure effectively, enabling fast and accurate inference across large MRI volumes.

The regression analyses on both simulated and patient data demonstrated that the model achieved low reconstruction error while maintaining strong cosine similarity to the spectrum. Through rigorous Leave-One-Out Cross-Validation, we observed consistent performance across patients with low variance (STD 0.62% for MAPE), confirming the model's robustness against inter-subject variability. While we acknowledge that the sample size of six patients is limited, this preliminary study demonstrates the feasibility of the framework and future validation on larger cohorts will be conducted to confirm clinical generalizability.

In conclusion, this work highlights the potential of masked transformer models to accelerate CEST-MRI by reducing the need for fully sampled acquisitions without compromising quantitative accuracy. By functioning as a spectral super-resolution module, the model allows us to acquire MR images with higher step size thus decreasing the acquisition time. Future works will focus on more enhancing model generalizability and incorporating spatial information through development of 2D vision transformers and 3D spatial-spectral transformers to further evaluate their reconstruction performance.

## ACKNOWLEDGMENTS

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA288379 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under Award Number RP240289 and RP240542. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## REFERENCES

- [1] Van Zijl, P. C. M., & Yadav, N. N. (2011). Chemical exchange saturation transfer (CEST): What is in a name and what isn't? *Magnetic Resonance in Medicine*, 65(4), 927–948. <https://doi.org/10.1002/mrm.22761>
- [2] Okuchi, S., Hammam, A., Golay, X., Kim, M., & Thust, S. (2020). Endogenous Chemical Exchange Saturation Transfer MRI for the diagnosis and therapy Response Assessment of brain tumors: A Systematic review. *Radiology Imaging Cancer*, 2(1), e190036. <https://doi.org/10.1148/rycan.2020190036>
- [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, 30, 5998–6008. <https://arxiv.org/pdf/1706.03762v5>
- [4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2010.11929>
- [5] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1711.05101>
- [6] Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/tip.2003.819861>