SPATIALLY ADAPTIVE CONVOLUTIONS VIA ATTENTION-ROUTED EXPERT MIXTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Convolutional autoencoders have been widely used for image reconstruction due to the local receptive fields and inductive biases of convolutional operators. However, conventional convolution applies the same learned transformation across spatial locations, limiting the ability of a single operator to adapt to heterogeneous local structures. We investigate whether this limitation can be addressed by learning a spatially varying composition of distinct convolutional transformations. We propose a mixture-of-experts (MoE) based adaptive convolutional autoencoder in which a fixed bank of structurally heterogeneous three-dimensional convolutional experts provides distinct transformations, while a spatial routing mechanism determines their contributions at different locations. In the encoder, localized window representations are processed by an attention module to produce spatially varying expert weights, which are interpolated to form a dense routing field. The resulting formulation retains the parameters of the individual experts while allowing their effective composition to vary spatially. We evaluate the framework on three-dimensional magnetic resonance imaging (MRI) reconstruction using the IXI-T1 dataset, focusing on the behavior of the learned experts and their spatial routing. The individual experts exhibit systematically different responses to image structure and spatial frequency, while routing redistributes their contributions across the feature map and retains a soft, non-uniform mixture rather than hard expert assignment. These observations are consistent with functional specialization emerging among heterogeneous convolutional operators through spatial routing.

1 INTRODUCTION

Convolutional neural networks (CNNs) have been a foundational architecture for visual representation learning, owing to the ability of convolutional operators to efficiently aggregate local spatial information (He et al., 2015; Singh et al., 2022). A conventional convolution applies shared learned kernels across the spatial extent of a feature map, providing an effective and parameter-efficient inductive bias that includes translation equivariance. However, this spatial weight sharing does not explicitly account for variations in local structure across different spatial regions (Cohen & Welling, 2016; Dai et al., 2017; Su et al., 2019).

This limitation has motivated extensive work on adaptive feature and convolutional operators, in which the effective transformation is conditioned on the input or its features (Hu et al., 2019; Brabandere et al., 2016; Yang et al., 2020), as well as spatially adaptive operators that allow the transformation to vary across spatial locations (Su et al., 2019; Li et al., 2021; Chandran et al., 2021; Chen et al., 2022). Dynamic and conditional convolution methods learn combinations of multiple kernels conditioned on the input, while other approaches adapt convolutional parameters, receptive fields, or sampling locations (Yang et al., 2020; Chen et al., 2020; Li et al., 2022). These approaches demonstrate ways of relaxing the fixed nature of convolution, but primarily do so by modifying or modulating the convolutional operator itself.

A complementary way to address this limitation is provided by Mixture-of-Experts (MoE), which combines a collection of learned expert transformations with a learned routing mechanism that determines their contributions (Jacobs et al., 1991; Shazeer et al., 2017). This allows different regions of a representation to use different combinations of learned transformations rather than being constrained to a single shared transformation (Dryden & Hoefler, 2022). From this perspective, MoE

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

offers a complementary form of adaptation: instead of modifying a single convolutional operator, multiple transformations can be made available while their relative contributions vary according to the input and spatial context.

We explore this idea using a fixed bank of structurally heterogeneous three-dimensional convolutional operators, each defined by a distinct local transformation. An attention-based spatial router determines the contribution of each expert across the feature map, allowing different combinations of transformations to be applied at different locations. This formulation provides a controlled setting for investigating whether spatial routing is associated with differentiated behavior among heterogeneous convolutional experts, and how their responses and routed contributions relate to image structure and frequency characteristics. We evaluate this framework on three-dimensional MRI reconstruction using the IXI-T1 dataset, focusing our analysis on the behavior of the learned experts and their spatial routing patterns.

Our contributions are summarized as follows:

- We introduce a spatially adaptive convolutional formulation consisting of structurally heterogeneous 3D convolutional experts with fixed parameters and spatially varying attention-based routing.
- We characterize the responses of the heterogeneous experts and their routed contributions with respect to image structure and frequency characteristics, providing evidence of differentiated expert behavior consistent with functional specialization.
- We analyze the resulting spatial routing behavior and expert composition, showing how the router forms a soft, non-uniform mixture of transformations across the feature map, and evaluate the formulation on 3D MRI reconstruction against uniform expert aggregation and alternative autoencoder architectures.

2 Related Work

Adaptive Convolution. Conventional convolution applies the same learned kernel across all spatial locations. This fixed transformation has motivated approaches that make the effective transformation adaptive to the input, with different methods introducing adaptation at different levels of the representation or convolutional operator. One early form of feature-level adaptation is Squeeze-and-Excitation Networks (Hu et al., 2019), which recalibrate feature channels according to the input. Dynamic Convolution (Chen et al., 2020) instead learns a bank of convolutional kernels and uses attention to compute input-dependent weights for combining them. KAConv (Shan et al., 2022) extends this idea by introducing spatially varying attention over kernel components. ODConv (Li et al., 2022) further generalizes dynamic convolution by introducing attention along multiple dimensions of the convolutional kernel, including the number of kernels, input channels, output channels, and spatial kernel dimensions. GLAU (Zheng et al., 2024) combines global channel attention with local spatial attention to generate convolutional kernels that adapt to both channel and spatial context. Other approaches have considered adaptation through the effective receptive field (Lee et al., 2019) or through learned sampling geometry (Dagès et al., 2025). Together, these methods demonstrate that the fixed convolutional operator can be made more flexible by adapting its parameters, receptive field, or sampling structure according to the input.

Mixture-of-Experts and Spatial Routing. Mixture-of-Experts (MoE) provides a complementary approach to adaptive computation by representing a transformation using multiple expert functions and a learned routing mechanism. In vision, MoE methods have explored different routing granularities, including token-level and spatially varying routing, allowing expert contributions to differ across regions of an input (Riquelme et al., 2021; Dryden & Hoefler, 2022). This spatial perspective is particularly relevant for visual data containing heterogeneous structures, where different regions may benefit from different combinations of transformations rather than sharing a single expert assignment (Han et al., 2025). MoE has also been incorporated directly into convolutional architectures, where convolutional transformations serve as the expert functions within CNNs. For example, ConvMoE (Zhang et al., 2023) uses convolutional layers as experts, establishing a direct connection between expert-based adaptive computation and convolutional representations.

These two directions provide complementary forms of adaptation. Adaptive convolution methods modify or modulate the convolutional operator itself, whereas MoE methods retain multiple transformations as distinct experts and learn how their contributions are combined. We investigate the latter form of adaptation using a fixed bank of structurally heterogeneous three-dimensional convolutional operators and spatially varying attention-based routing, with particular emphasis on the resulting expert responses and their functional differentiation.

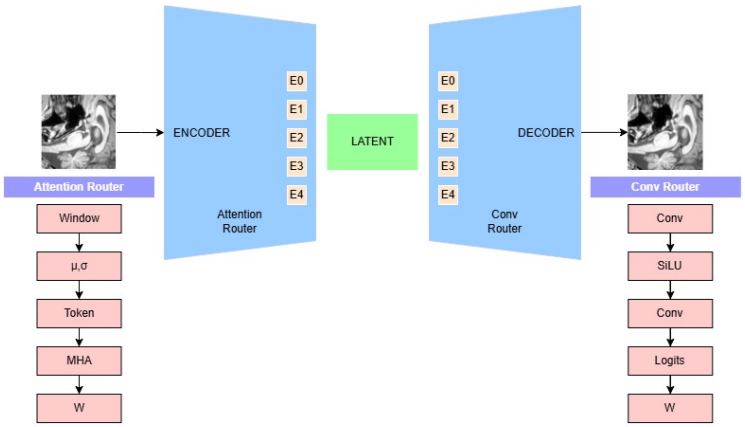


Figure 1: Architectural Overview of the Proposed Spatially Adaptive Convolution Architecture

3 METHODS

We formulate spatially adaptive convolution using a fixed bank of structurally heterogeneous 3D convolutional operators and a learned router that determines their contribution at different spatial locations.

Adaptive convolutions serve two complementary roles in our autoencoder. In the encoder, spatially adaptive mixtures of convolutional experts transform the image into a latent representation. While the decoder adapts the transformations to reconstruct the image from the latent. The encoder and decoder each contains three convolutional blocks. To limit computational overhead from adaptive routing, we apply routing in the deeper encoder blocks and the first decoder block. A formal derivation of the proposed spatially adaptive convolution formulation is provided in Appendix A.

Table 1: Structural characteristics of the encoder and decoder experts.

Expert	**Encoder**	**Decoder**	**Characteristics**
E_0	Low-pass $1 \times 3 \times 3 \rightarrow 1 \times 1 \times 1$	$1 \times 1 \times 1$ on low-pass component	Low-frequency / smoothing
E_1	$1 \times 1 \times 1$	$1 \times 1 \times 1$ on high-pass residual	High-frequency-conditioned
E_2	$1 \times 3 \times 3$	$1 \times 3 \times 3$	In-plane spatial mixing
E_3	$3 \times 1 \times 1$	$3 \times 1 \times 1$	Through-plane
E_4	$1 \times 1 \times 1$	$1 \times 1 \times 1$	Pointwise

3.1 ENCODER

We divide our encoder into two parts, early encoder and deep encoder each using two encoder encoder blocks, with spatial downsampling between consecutive blocks. Each downsampling block reduces the in-plane spatial dimensions according to $D, H, W \rightarrow D, H/2, W/2$. while preserving the through-plane dimension. Each encoder block contains five heterogeneous convolutional experts $E0, \ldots E4$. Table 1 shows the convolutions used in each expert and its charcateristics. The early encoders use the experts, but do not perform routing.

We introduce spatial routing in the deep encoder. Where the expert contributions are weighted according to the learned spatially varying coefficients.

3.1.1 Routing Processing

To construct the input to the routing network, we first normalize the feature representation spatially on a per-channel basis. We additionally extract a high-frequency component using a local average filter, $x_{\mathrm{HF}} = x - \mathrm{AvgPool}_{3\times3\times3}(x)$, and add a scaled version of this component to the normalized representation. This provides the routing network with information about local feature variation in addition to the underlying feature representation.

The resulting representation is downsampled in the in-plane dimensions by a factor of two and projected to a reduced number of channels. A second high-frequency component is then computed from this reduced representation and combined with its low-frequency component. Specifically, we use $x_{\mathrm{HF}} = x_{\mathrm{low}} - \mathrm{AvgPool}_{3\times3\times3(x_{\mathrm{low}})}$ and construct the routing representation as $x_{\mathrm{route}} = x_{\mathrm{low}} + 0.5x_{\mathrm{HF}}$. The coefficient 0.5 is used as a heuristic scaling factor. This preprocessed representation is subsequently partitioned into spatial windows for token construction and attention-based routing.

3.1.2 Window Tokenization and Attention

We partition the feature representation into non-overlapping three-dimensional spatial windows. For each window we summarize the feature activations using mean(μ_q) and standard deviation(σ_q) of the across the channel dimension. The mean captures the aggregate activation of the window, while the standard deviation provides a measure of the variation within the window. These statistics provide a compact representation of both the features content and its variations for determining the routing coefficients. We construct the corresponding window token as $t_q = \mu_q + 0.1\sigma_q$,. Where we use 0.1 as a heuristic scaling factor to incorporate within-window variation without having the standard deviation term dominate the mean. The resulting tokens are normalized across their feature dimensions and processed using multi-head self-attention.

Given the set of window tokens T, the attention operation is written as $H = \mathrm{MHA}(T, T, T)$. The contextualized representation H is then projected to five dimensions using a learned linear projection, $L_q = W_{\mathrm{route}}H_q + b_{\mathrm{route}}$, where each dimension corresponds to one convolutional expert. Thus, the attention mechanism produces five routing logits for each spatial window, which are subsequently transformed into spatially varying expert weights.

3.1.3 Spatial Routing and Expert Aggregation

The attention module produces five routing logits for each localized three-dimensional spatial window. These window-level logits are reshaped according to their corresponding three-dimensional window grid and interpolated to the spatial resolution of the feature representation, producing a dense routing field $L \in \mathbb{R}^{5\times D\times H\times W}$. Thus, routing decisions are derived from localized spatial windows rather than predicted independently at every voxel, while interpolation allows the resulting routing coefficients to vary continuously across the feature map.

Before computing the routing coefficients, the logits are normalized to control their spatial and expert-wise variation. The normalized logits are then converted into routing coefficients using a temperature-scaled softmax,

$$w_k(p) = \frac{\exp(L_k(p)/\tau)}{\sum_{j=0}^{4}\exp(L_j(p)/\tau)}, \tag{1}$$

where $\tau = 0.8$ and p denotes a spatial location. Consequently, the routing coefficients satisfy $\sum_{k=0}^{4} w_k(p) = 1$ at every spatial location.

The five convolutional experts independently transform the input feature representation to produce $F_k = f_k(x)$. The final output of the adaptive encoder block is obtained by taking a spatially varying weighted combination of these expert outputs,

$$y(p) = \sum_{k=0}^{4} w_k(p)F_k(p). \tag{2}$$

The resulting representation is subsequently processed by Group Normalization and a SiLU activation. Unlike conventional convolution, where the same transformation is applied at every spatial location, the proposed formulation allows the effective convolutional transformation to vary spatially through the learned routing field while keeping the parameters and structures of the individual experts fixed.

3.2 Decoder

The decoder progressively transforms the latent representation back to the spatial resolution of the input. It consists of three convolutional blocks, with spatial upsampling between the first two blocks. As in the encoder, each block contains five heterogeneous convolutional experts. Adaptive routing is applied only in the first decoder block, while the remaining decoder blocks use a uniform mixture of the five experts.

The first decoder block receives the latent representation and reduces the number of feature channels before applying the expert transformations. For subsequent decoder blocks, the same heterogeneous expert bank is used without adaptive routing. Spatial upsampling is performed only in the in-plane dimensions, such that $D, H, W \to D, 2H, 2W$.

3.2.1 Frequency Decomposition and Expert Transformations

Before applying the decoder expert bank, we decompose the input feature representation into low- and high-frequency components using a local in-plane average filter, $x_{\text{low}} = \text{AvgPool}_{1\times3\times3}(x)$, and $x_{\text{high}} = x - x_{\text{low}}$. The low- and high-frequency components are provided to separate pointwise experts, while the remaining experts operate directly on the feature representation. Table 1 shows the decoder experts used and their characteristics.

3.2.2 Adaptive Decoder Routing

In the first decoder block, the expert contributions are combined using spatially varying routing coefficients. Unlike the attention-based routing used in the deep encoder, the decoder routing coefficients are predicted directly from the feature representation using a lightweight convolutional routing network. The router consists of a $3 \times 3 \times 3$ convolution followed by a SiLU activation and a $1 \times 1 \times 1$ convolution that produces five routing logits at each spatial location.

The logits are converted into routing coefficients using a softmax,

$$w_k(p) = \frac{\exp(L_k(p))}{\sum_{j=0}^{4} \exp(L_j(p))}. \tag{3}$$

The expert outputs are then combined according to $y(p) = \sum_{k=0}^{4} w_k(p) F_k(p)$. For the remaining decoder blocks, the routing coefficients are fixed to $1/5$ for all experts and spatial locations.

3.2.3 Upsampling and Output Refining

Following the first decoder block, the feature representation is upsampled by a factor of two in the in-plane dimensions while preserving the through-plane dimension. The subsequent decoder blocks further transform the upsampled representation using the uniform heterogeneous expert mixture. Finally, an anisotropic output refinement head combines $1 \times 3 \times 3$ and $3 \times 1 \times 1$ convolutions to produce the single-channel reconstructed volume.

4 Experiments

We focus our analysis on whether spatial routing is associated with differentiated behavior among the heterogeneous convolutional experts. We examine the learned expert responses and routed contributions with respect to observable properties of the input, including image structure and frequency characteristics. We focus primarily on the decoder routing, whose operation closer to the image space provides a more direct basis for relating expert contributions to observable image structures. To additionally characterize the behavior of the encoder routing, we provide visualizations of its spatially varying expert contributions in AppendixC.2.

We evaluate the proposed formulation on the IXI-T1 (ixi, 2026) dataset using an 80-10-10 train-validation-test split of 3D volumes. As a reconstruction testbed, we compare the proposed model with the MONAI framework's autoencoder (Cardoso et al., 2022) and a three-dimensional variant of the standard ViT (Dosovitskiy et al., 2021). We first characterize the responses of the individual experts and their routed contributions, examining their relationships with image structure and frequency. We then examine the behavior of spatial routing through comparisons between adaptive and uniform expert aggregation, followed by expert-removal and reduced-expert ablations. All models are trained for 20 epochs under the same reconstruction objective.

4.1 Spatial Routing of Experts

We first characterize the spatial behavior of the learned routing mechanism by visualizing the expert contributions and measuring their distribution across the test set. Specifically, we examine the mean and standard deviation of the routing weights for each expert, as well as the fraction of spatial locations at which each expert receives the highest routing weight. These measurements characterize whether the learned routing is spatially uniform or varies across inputs and locations, and whether the resulting mixture is well described by hard expert assignment.

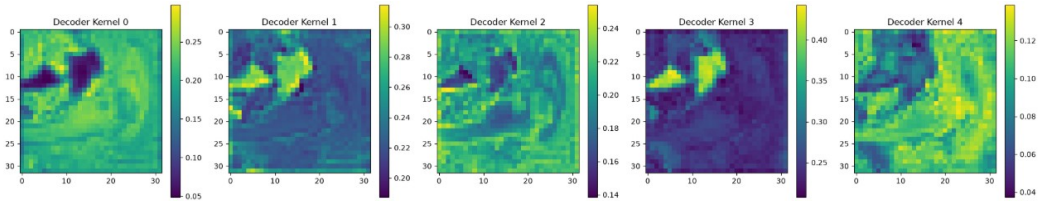
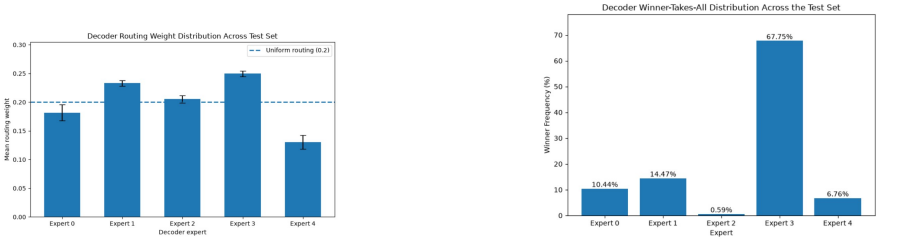


Figure 2: Expert Routing Visualization



(a) Routing contributions of different experts.

(b) Winner-takes-all expert distribution.

Figure 3: Analysis of expert routing behavior.

Results. Figure 2 visualizes the spatially varying routing weights of the five experts for a representative test volume. The routing patterns differ across experts and vary spatially within the volume, indicating that the relative contribution of the experts is not determined by a fixed global weighting. Additional visualizations of the encoder and decoder routing maps across representative test volumes are provided in Appendix C.2.

Figure 3a reports the mean and standard deviation of the routing weights across the test dataset. The non-zero standard deviations indicate that the contribution assigned to each expert varies across test volumes. Figure 3b further reports Winner-Takes-All (WTA) dominance, defined as the fraction of spatial locations for which each expert receives the highest routing weight. For example, Expert 2 receives approximately 20% of the mean routing weight but is the dominant expert at only 0.59% of spatial locations. Conversely, another expert is the largest contributor at approximately 68% of locations. This does not imply hard expert assignment, since WTA dominance considers only the identity of the largest weight at each location, whereas the remaining experts may still re-

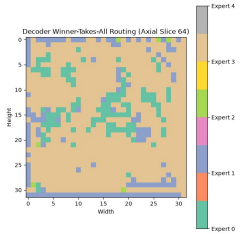


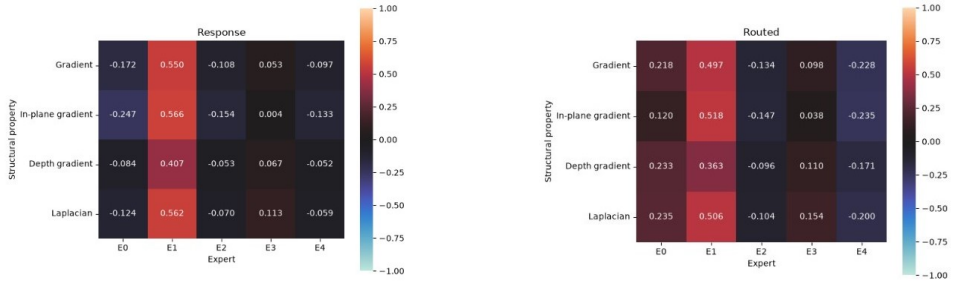
Figure 4: Winner-Takes-All visualization

ceive substantial weights. This is consistent with the high normalized routing entropy reported in Table 2, indicating that the overall routing distribution remains soft. Thus, WTA dominance and routing entropy capture complementary aspects of the routing behavior: the former identifies the largest contributor, while the latter measures how distributed the full mixture remains.

Figure 4 visualizes the spatial distribution of the dominant experts, further showing that the identity of the expert with the largest routing weight varies across spatial regions of the volume. Together, these observations show that the routing mechanism forms a spatially varying, soft composition of the available convolutional experts rather than assigning each spatial location to a single expert.

4.2 Structural and Frequency Dependence

We characterize the relationship between expert responses, spatial routing, and observable characteristics of the input through structural and frequency analyses. We first compute structural maps from the input, including gradient magnitude, in-plane gradient magnitude, through-plane gradient magnitude, and Laplacian magnitude. We then measure the correlation between these structural quantities and the response magnitude of each expert, both before and after spatial routing. We additionally characterize the frequency response of the individual experts using controlled sinusoidal inputs with varying spatial frequency along a fixed in-plane direction.



(a) Correlations before Routing

(b) Correlations after Routing

Figure 5: Heatmaps of correlations between experts and different structural highlighters before and after routing

Results. Figure 5a reports the correlations between the unrouted expert responses and the structural measures, while Figure 5b shows the corresponding correlations for the routed contributions. For an expert k, we compute its response magnitude as

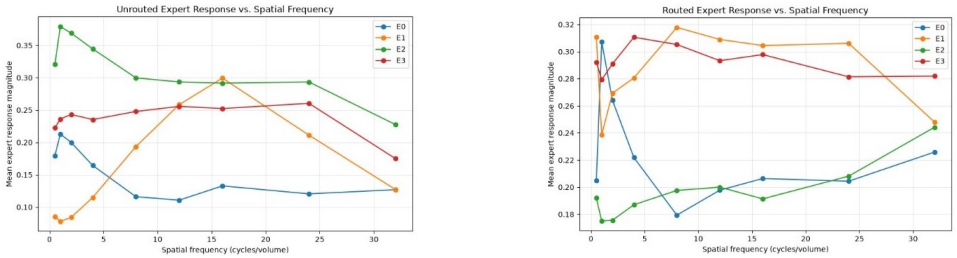
$$A_k(p) = \sqrt{\frac{1}{C}\sum_{c=1}^{C} F_{k,c}(p)^2}, \tag{4}$$

where F_k denotes the output of the corresponding expert. We compare the correlation between A_k and each structural map with the correlation obtained after weighting the response by the spatial routing coefficient, $w_k(p)A_k(p)$.

The unrouted responses exhibit distinct correlation patterns across the experts. In particular, Expert 1 exhibits the strongest positive correlations across the four structural measures. Following spatial routing, these correlation patterns change substantially. For example, Expert 0 changes from negative correlations across the four measures to positive correlations, while Expert 4 becomes more negatively associated with the structural measures. Expert 1 remains positively associated with the structural quantities, although the magnitude of its correlations is reduced relative to the unrouted response. These observations show that the spatial routing weights alter the observed structural associations of the expert contributions rather than simply preserving the response patterns of the underlying operators.

Figures 6a and 6b further characterize the frequency-dependent behavior of the experts using controlled sinusoidal inputs. The unrouted experts exhibit distinct frequency response profiles, with differences in both response magnitude and variation across spatial frequencies. After spatial routing, the resulting expert contributions exhibit different frequency-dependent profiles, with the relative contributions becoming more similar over parts of the frequency range. Thus, the individual

operators exhibit systematically different frequency responses, while spatial routing modulates their relative contributions as the input frequency changes.



(a) Frequency Raw Responses

(b) Frequency Routed Responses

Figure 6: Frequency characteristics of the raw and routed learned experts.

Together, the structural and frequency analyses show that the heterogeneous experts exhibit differentiated response characteristics and that spatial routing changes their relative contributions in relation to observable properties of the input. These results provide evidence consistent with functional specialization among the heterogeneous convolutional experts.

4.3 Ablation studies

We assess the role of spatial routing and expert composition through a series of controlled ablation studies. We compare the full five-expert configuration with reduced expert banks to examine how changing the available expert composition affects the similarity of expert responses and the resulting routing behavior. For each configuration, we measure the mean absolute pairwise correlation between expert response maps as a measure of response redundancy, together with the normalized routing entropy and mean routing usage of the individual experts.

Additional ablation and characterization experiments are provided in the appendix C. These include comparisons of encoder- and decoder-side routing, structural correlations between expert responses and input properties, frequency-response analyses of the individual experts and their routed contributions, and reconstruction metrics across different expert configurations. These analyses provide additional characterization of the learned expert composition beyond the response redundancy and routing statistics presented here.

Results. Table 2 summarizes the response redundancy, normalized routing entropy, and expert usage across the different expert configurations. Changing the composition of the expert bank substantially alters both the similarity of expert responses and their routing distributions. In particular, removing E_1 increases the mean absolute response correlation from 0.60 to 0.88, indicating substantially greater similarity among the remaining expert responses. In contrast, the E_0, E_3, E_4 configuration exhibits a much lower response correlation of 0.08, demonstrating that different expert compositions can produce markedly different levels of response redundancy.

Table 2: Expert composition ablation based on response redundancy, normalized routing entropy, and mean expert usage. A dash indicates that the corresponding expert is removed from the configuration.

Configuration	Response Redundancy	Norm. Entropy	E0	E1	E2	E3	E4
5E	0.6027	0.9573	0.182	0.233	0.205	0.250	0.130
$-E_0$	0.5542	0.9542	–	0.288	0.220	0.282	0.210
$-E_1$	0.8854	0.9688	0.239	–	0.309	0.209	0.243
$-E_2$	0.5517	0.9599	0.261	0.323	–	0.170	0.246
$-E_3$	0.5215	0.979	0.207	0.321	0.229	–	0.2424
$-E_4$	0.5259	0.9627	0.193	0.315	0.197	0.295	–
$-(E_0, E_4)$	0.4423	0.9808	–	0.311	0.296	0.393	–
$-(E_0, E_3, E_4)$	0.0781	0.9593	–	0.566	0.434	–	–

The routing statistics further show that the full five-expert model retains a soft mixture of experts, with a normalized routing entropy of 0.957. Expert usage is nevertheless non-uniform, ranging from 0.130 for E_4 to 0.250 for E_3, while the spatial variance of the routing weights indicates that the contribution of individual experts also varies across the feature map. The reduced expert configurations exhibit corresponding changes in their routing distributions. For example, removing E_0 substantially increases the spatial variance of the routing weight associated with E_3, while removing E_1 produces a higher normalized routing entropy of 0.969.

Together, these results show that changing the available expert set reorganizes both the response similarity of the remaining experts and the distribution of their spatial contributions. Thus, the learned expert composition depends on the set of operators available to the model, while retaining the overall behavior of a soft, spatially varying mixture of experts.

5 Conclusion

We presented a spatially adaptive convolutional formulation that combines a fixed bank of structurally heterogeneous three-dimensional convolutional operators through spatially varying routing. Rather than modifying the parameters of individual convolutional kernels, the proposed formulation retains fixed expert transformations while adapting their relative contributions across the spatial extent of the feature map.

We evaluated the formulation on three-dimensional MRI reconstruction and used this setting to characterize the behavior of the resulting experts. The individual experts exhibit systematically different response characteristics with respect to image structure and spatial frequency, while the routing mechanism produces a non-uniform, spatially varying mixture of their contributions. Structural correlation and frequency analyses further show that routing changes the observed associations between individual expert contributions and characteristics of the input. Expert-composition ablations additionally demonstrate that changing the available operator set reorganizes both response similarity and routing behavior. Together, these results provide evidence of differentiated expert behavior that is consistent with functional specialization among heterogeneous convolutional operators, while retaining a soft, compositional mixture rather than enforcing hard expert assignment.

These results suggest that spatial adaptation can be achieved by controlling the composition of fixed convolutional transformations, providing an alternative to approaches that directly modify convolutional parameters. Future work could investigate more expressive routing mechanisms and alternative expert designs, as well as whether the observed expert differentiation provides useful representations for downstream medical imaging tasks such as classification and segmentation. Further investigation could also explore the application of the proposed formulation to other three-dimensional imaging tasks and architectures.

AI use statement

In this work, we used generative AI tools for research ideation and hypothesis refinement, development and refinement of the conceptual framework, formulation and development of mathematical formulations and proofs, design and analysis of experiments, implementation and debugging of methods, and interpretation of experimental results. We also used generative AI tools for literature search and analysis, brainstorming, suggesting experimental analyses and parameters, drafting and editing parts of the manuscript, improving readability and organization, and assisting with software code.

We did not use generative AI tools to generate experimental data or experimental results. All AI-assisted mathematical formulations and proofs were reviewed and independently verified by the authors. AI-assisted code was tested against the actual experimental implementations, and AI-assisted literature suggestions and claims were independently checked against the relevant sources. AI-assisted manuscript text was reviewed and revised by the authors. Tasks involving translation, qualitative or thematic data analysis, and synthetic dataset generation were not applicable to this work. The authors take full responsibility for the final content of this work, including all text, claims, analyses, code, mathematical formulations, proofs, and other artifacts produced with the aid of generative AI.

Ethics statement

Our experiments use the publicly available IXI-T1 dataset, which contains MRI scans collected from healthy subjects at three hospitals in London as part of the IXI (Information eXtraction from Images) project. The dataset is publicly released under the Creative Commons Attribution-ShareAlike 3.0 license. The specific image files used in our experiments contain only the MRI volumes and do not include demographic or clinical metadata. No new human-subject data were collected for this study, and the authors did not recruit or interact with participants. We used the publicly released data solely for computational research in accordance with the applicable terms of use. The IXI dataset documentation does not provide a specific institutional ethics approval number or detailed participant-consent information; accordingly, we do not make additional claims regarding the ethics approval or consent procedures of the original data collection. No clinical decisions, diagnoses, or interventions are made using the resulting models, and the experiments are not intended to provide clinical predictions for individual subjects.

Reproducibility statement

Detailed information on the model architecture, expert configurations, routing mechanism, data preprocessing, training procedure, evaluation protocol, and experimental settings is provided in appendix D. All experiments were conducted on an NVIDIA RTX A5000 GPU. Random seeds are controlled for experiments involving multiple initializations, and seed-level variability is reported where applicable. Source code, configuration files, and scripts required to reproduce the reported experiments will be made publicly available upon publication.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

IXI Dataset: Information extraction from images. https://brain-development.org, 2026. RRID:SCR$_0$05839.

Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic Filter Networks, June 2016. URL http://arxiv.org/abs/1605.09673. arXiv:1605.09673 [cs.LG].

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. MONAI: An open-source framework for deep learning in healthcare, November 2022. URL http://arxiv.org/abs/2211.02701. arXiv:2211.02701 [cs.LG].

Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive Convolutions for Structure-Aware Style Transfer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7968–7977, June 2021. doi: 10.1109/CVPR46437.2021.00788. URL https://ieeexplore.ieee.org/document/9578617. ISSN: 2575-7075.

Jierun Chen, Tianlang He, Weipeng Zhuo, Li Ma, Sangtae Ha, and S.-H. Gary Chan. TVConv: Efficient Translation Variant Convolution for Layout-aware Visual Processing, March 2022. URL http://arxiv.org/abs/2203.10489. arXiv:2203.10489 [cs.CV].

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic Convolution: Attention over Convolution Kernels, March 2020. URL http://arxiv.org/abs/1912.03458. arXiv:1912.03458 [cs.CV].

Taco S. Cohen and Max Welling. Group Equivariant Convolutional Networks, June 2016. URL http://arxiv.org/abs/1602.07576. arXiv:1602.07576 [cs.LG].

Thomas Dagès, Michael Lindenbaum, and Alfred M. Bruckstein. Metric Convolutions: A Unifying Theory to Adaptive Image Convolutions, July 2025. URL http://arxiv.org/abs/2406.05400. arXiv:2406.05400 [cs.CV].

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable Convolutional Networks, June 2017. URL http://arxiv.org/abs/1703.06211. arXiv:1703.06211 [cs.CV].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs.CV].

Nikoli Dryden and Torsten Hoefler. Spatial mixture-of-experts. *Advances in Neural Information Processing Systems*, 35:11697–11713, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4c5e2bcbf21bdf40d75fddad0bd43dc9-Abstract-Conference.html.

Xumeng Han, Longhui Wei, Zhiyang Dou, Yingfei Sun, Zhenjun Han, and Qi Tian. ViMoE: An Empirical Study of Designing Vision Mixture-of-Experts. *IEEE Trans. on Image Process.*, 34: 7209–7221, 2025. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2025.3626887. URL https://ieeexplore.ieee.org/document/11226953/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs.CV].

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks, May 2019. URL http://arxiv.org/abs/1709.01507. arXiv:1709.01507 [cs.CV].

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, February 1991. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1991.3.1.79. URL https://direct.mit.edu/neco/article/3/1/79-87/5560.

Yekang Lee, Heechul Jung, Dongyoon Han, Kyungsu Kim, and Junmo Kim. Learning receptive field size by learning filter size. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1203–1212. IEEE, 2019. URL https://ieeexplore.ieee.org/abstract/document/8658949/.

Chao Li, Aojun Zhou, and Anbang Yao. Omni-Dimensional Dynamic Convolution, September 2022. URL http://arxiv.org/abs/2209.07947. arXiv:2209.07947 [cs.CV].

Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the Inherence of Convolution for Visual Recognition, March 2021. URL https://arxiv.org/abs/2103.06255v2.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling Vision with Sparse Mixture of Experts, June 2021. URL https://arxiv.org/abs/2106.05974v1.

Xinxin Shan, Tai Ma, Yutao Shen, Jiafeng Li, and Ying Wen. KAConv: Kernel attention convolutions. *Neurocomputing*, 514:477–485, December 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.10.017. URL https://www.sciencedirect.com/science/article/pii/S0925231222012796.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer, January 2017. URL http://arxiv.org/abs/1701.06538. arXiv:1701.06538 [cs.LG].

Pravendra Singh, Pratik Mazumder, and Vinay P. Namboodiri. Context extraction module for deep convolutional neural networks. *Pattern Recognition*, 122:108284, February 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108284. URL https://www.sciencedirect.com/science/article/pii/S0031320321004647.

Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-Adaptive Convolutional Neural Networks, April 2019. URL http://arxiv.org/abs/1904.05373. arXiv:1904.05373 [cs.CV].

Brandon Yang, Gabriel Bender, Quoc V. Le, and Jiquan Ngiam. CondConv: Conditionally Parameterized Convolutions for Efficient Inference, September 2020. URL http://arxiv.org/abs/1904.04971. arXiv:1904.04971 [cs.CV].

Yihua Zhang, Ruisi Cai, Tianlong Chen, Guanhua Zhang, Huan Zhang, Pin-Yu Chen, Shiyu Chang, Zhangyang Wang, and Sijia Liu. Robust Mixture-of-Expert Training for Convolutional Neural Networks, August 2023. URL http://arxiv.org/abs/2308.10110. arXiv:2308.10110 [cs.CV].

Chusan Zheng, Yafeng Li, Jian Li, Ning Li, Pan Fan, Jieqi Sun, and Penghui Liu. Dynamic convolution neural networks with both global and local attention for image classification. *Mathematics*, 12(12): 1856, 2024. URL https://www.mdpi.com/2227-7390/12/12/1856.

A Mathematical Formulation

This appendix provides the mathematical formulation of the proposed spatially adaptive convolution, including the heterogeneous expert aggregation, attention-based encoder routing, convolutional decoder routing, and the structural analysis used to characterize expert contributions.

A.1 Spatially Adaptive Expert Convolution

Let $X \in \mathbb{R}^{C\times D\times H\times W}$ denote an input feature map and let $\mathcal{F} = \{f_k\}_{k=1}^{K}$ denote a bank of K heterogeneous convolutional experts. Each expert produces an output feature map

$$F_k(p) = f_k(X)(p), \tag{5}$$

where p denotes a spatial location. A spatial router produces a routing vector

$$\mathbf{w}(p) = [w_1(p), \ldots, w_K(p)], \tag{6}$$

subject to

$$w_k(p) \geq 0, \qquad \sum_{k=1}^{K} w_k(p) = 1. \tag{7}$$

The output of the spatially adaptive convolution is given by

$$\boxed{Y(p) = \sum_{k=1}^{K} w_k(p) F_k(p)} \tag{8}$$

where the expert parameters remain fixed after training and the spatial adaptation is achieved through the spatially varying coefficients $w_k(p)$.

For linear convolutional experts, an individual expert can be written as

$$F_k(p) = \sum_{\delta \in \Omega_k} K_k(\delta) X(p+\delta), \tag{9}$$

where K_k denotes the kernel of expert k and Ω_k its receptive field. Substituting this expression into Eq. 8 gives

$$Y(p) = \sum_k w_k(p) \sum_{\delta \in \Omega_k} K_k(\delta) X(p+\delta). \tag{10}$$

This can be expressed using an effective spatially varying kernel as

$$Y(p) = \sum_\delta K_{\text{eff}}(p,\delta) X(p+\delta), \tag{11}$$

where

$$\boxed{K_{\text{eff}}(p,\delta) = \sum_k w_k(p) K_k(\delta)} \tag{12}$$

Thus, the effective transformation can vary with spatial position through the routing weights, while the individual expert kernels K_k remain fixed.

A.2 Attention-Based Encoder Routing

The encoder routing mechanism derives spatially varying expert weights from localized feature windows. Given an input feature map X, the feature map is partitioned into non-overlapping three-dimensional windows. For a spatial position r within window q, the channel-wise mean and standard deviation are computed as

$$\mu_q(r) = \frac{1}{C} \sum_{c=1}^{C} X_q(r,c), \tag{13}$$

$$\sigma_q(r) = \sqrt{\frac{1}{C} \sum_{c=1}^{C} \left(X_q(r,c) - \mu_q(r)\right)^2}. \tag{14}$$

The statistics are combined to form a compact routing descriptor

$$t_q(r) = \mu_q(r) + \lambda_\sigma \sigma_q(r), \tag{15}$$

where $\lambda_\sigma = 0.1$. The resulting descriptors are normalized and arranged as a sequence of spatial tokens. Multi-head self-attention is then applied to model interactions between the localized windows.

For a token matrix T, the query, key, and value representations are

$$Q = TW_Q, \qquad K = TW_K, \qquad V = TW_V. \tag{16}$$

The attention operation is given by

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{17}$$

followed by

$$T_{\text{attn}} = AV. \tag{18}$$

The attended tokens are projected to K expert logits,

$$L_q = T_{\text{attn}} W_R + b_R, \qquad L_q \in \mathbb{R}^K, \tag{19}$$

where each window therefore produces one routing logit for each expert. The window-level logits are reshaped into their spatial window arrangement and interpolated to obtain dense spatial logits $L_k(p)$.

The logits are normalized across experts at each spatial location:

$$\tilde{L}_k(p) = \frac{L_k(p) - \mu_L(p)}{\sigma_L(p) + \epsilon}, \tag{20}$$

where

$$\mu_L(p) = \frac{1}{K} \sum_{k=1}^{K} L_k(p) \tag{21}$$

and

$$\sigma_L(p) = \text{Std}_k \left(L_k(p) \right). \tag{22}$$

During training, a small amount of Gaussian noise is added to the normalized logits. The final routing weights are obtained using a temperature-scaled softmax:

$$\boxed{w_k(p) = \frac{\exp\left(\tilde{L}_k(p)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\tilde{L}_j(p)/\tau\right)}} \tag{23}$$

where $\tau = 0.8$ is the routing temperature. The resulting weights form a dense spatial routing field that is used in Eq. 8 to combine the heterogeneous experts.

A.3 Convolutional Decoder Routing

The decoder employs a lightweight convolutional router to produce spatially varying expert weights directly from the decoder feature map. Given a decoder feature map X, the intermediate routing representation is computed as

$$H = \text{SiLU}\left(\text{Conv}_{3\times3\times3}(X) \right). \tag{24}$$

The intermediate representation is then projected to K spatial routing logits using a pointwise convolution:

$$L = \text{Conv}_{1\times1\times1}(H), \qquad L \in \mathbb{R}^{K\times D\times H\times W}. \tag{25}$$

The routing weights are obtained independently at each spatial location using a softmax over the expert dimension:

$$\boxed{w_k(p) = \frac{\exp(L_k(p))}{\sum_{j=1}^{K} \exp(L_j(p))}} \tag{26}$$

The decoder output is subsequently formed using the same spatial expert aggregation defined in Eq. 8.

The encoder and decoder therefore share the same expert-composition formulation while employing different routing mechanisms. The encoder uses attention over localized windows to derive its routing fields, whereas the decoder uses a lightweight convolutional router to directly predict dense spatial logits.

A.4 Uniform and Adaptive Expert Aggregation

To isolate the effect of spatial routing, we compare adaptive expert aggregation against a uniform aggregation of the same heterogeneous expert bank. For the uniform configuration, all experts receive equal weight at every spatial location:

$$\boxed{w_k(p) = \frac{1}{K}} \tag{27}$$

and therefore

$$Y_{\text{uniform}}(p) = \frac{1}{K} \sum_{k=1}^{K} F_k(p). \tag{28}$$

In contrast, the adaptive configuration uses spatially varying routing weights:

$$Y_{\text{adaptive}}(p) = \sum_{k=1}^{K} w_k(p) F_k(p). \tag{29}$$

Thus, the comparison between the uniform and adaptive configurations keeps the heterogeneous expert transformations fixed while changing only how their contributions are composed spatially.

A.5 Structural Analysis of Expert Contributions

To characterize the relationship between expert responses and image structure, we first compute the response magnitude of each expert. For expert k, this is defined as

$$A_k(p) = \sqrt{\frac{1}{C} \sum_{c=1}^{C} F_{k,c}(p)^2}. \tag{30}$$

The corresponding routed contribution magnitude is defined as

$$\boxed{C_k(p) = w_k(p) A_k(p)} \tag{31}$$

where $w_k(p)$ is the spatial routing weight. Note that the actual network output remains the signed expert aggregation in Eq. 8; Eq. 31 is used only for analyzing the magnitude of each expert's routed contribution.

Given a structural map $S(p)$, such as gradient magnitude, in-plane gradient magnitude, through-plane gradient magnitude, or Laplacian magnitude, we measure its linear association with each expert response using the Pearson correlation coefficient:

$$\rho_k^{\text{raw}} = \text{Corr}\left(A_k(p), S(p)\right), \tag{32}$$

and, for the routed contribution,

$$\boxed{\rho_k^{\text{routed}} = \text{Corr}\left(C_k(p), S(p)\right).} \tag{33}$$

These correlations are used to characterize associations between expert responses and structural properties of the input, rather than to infer a causal relationship between routing and image structure.

A.6 Routing Entropy

To quantify the concentration of the spatial routing distribution, we compute the entropy of the routing weights at each spatial location:

$$H(p) = -\sum_{k=1}^{K} w_k(p) \log w_k(p). \tag{34}$$

Since the maximum entropy of a distribution over K experts is $\log K$, we use the normalized routing entropy

$$\boxed{H_{\text{norm}}(p) = \frac{-\sum_{k=1}^{K} w_k(p) \log w_k(p)}{\log K}} \tag{35}$$

such that

$$0 \leq H_{\text{norm}}(p) \leq 1. \tag{36}$$

Values close to 1 indicate approximately uniform expert usage, whereas values close to 0 indicate concentrated routing toward a small number of experts. We use this measure to characterize the degree of spatial selectivity in the routing fields, including those produced by the encoder attention router.

B Implementation PseudoCode

In this appendix, algorithm 1 and algorithm 2 provide the implementation psuedocode for the encoder and decoder of our model respectively.

C Additional Ablation and Analysis

This appendix provides additional experiments and visualizations supporting the analysis presented in the main paper. We first report reconstruction results for the different expert configurations, followed by detailed analyses of expert composition, spatial routing, structural associations, and frequency-dependent behavior. We additionally provide qualitative visualizations of the encoder and decoder routing fields.

C.1 Reconstruction and Performance Metrics

We evaluate reconstruction quality using mean MSE, PSNR, and SSIM across the test dataset. To characterize the computational requirements of each model, we additionally report the number of trainable parameters, peak VRAM usage, and inference time.

Results Table 3 reports mean reconstruction metrics across the test dataset. Our model achieves higher reconstruction fidelity than both baseline models across the evaluated reconstruction metrics. Table 2 reports the corresponding model's performance characteristics. Our model uses 56.9% and 87.0% fewer parameters than the convolutional and attention-based baselines, respectively, while requiring higher peak VRAM and observing longer inference times.

Algorithm 1 Attention-Routed Encoder Block

Require: Input feature map $x \in \mathbb{R}^{C\times D\times H\times W}$
Ensure: Output feature map y

$x_{\text{HF}} \leftarrow x - \text{AvgPool}_{3\times3\times3}(x)$
$x_{\text{HF}} \leftarrow \text{clip}(x_{\text{HF}}, -3, 3)$
$x \leftarrow \text{Normalize}_{DHW}(x)$
$x_{\text{route}} \leftarrow x + 0.5x_{\text{HF}}$
$x_{\text{route}} \leftarrow \text{Downsample}_{H,W}(x_{\text{route}})$
$x_{\text{route}} \leftarrow \text{Center}_{DHW}(x_{\text{route}})$
$x_{\text{route}} \leftarrow \text{ReduceChannels}(x_{\text{route}})$
$x^r_{\text{HF}} \leftarrow x_{\text{route}} - \text{AvgPool}_{3\times3\times3}(x_{\text{route}})$
$x_{\text{low}} \leftarrow \text{Normalize}_{DHW}(x_{\text{route}})$
$x_{\text{high}} \leftarrow \text{Normalize}_{DHW}(x^r_{\text{HF}})$
$x_{\text{route}} \leftarrow x_{\text{low}} + 0.5x_{\text{high}}$
$T \leftarrow \text{WindowTokenize}(x_{\text{route}})$
$T \leftarrow \text{MHA}(T, T, T)$
$L \leftarrow \text{Linear}(T, K)$
$L \leftarrow \text{ReshapeToWindowGrid}(L)$
$L \leftarrow \text{InterpolateToFeatureResolution}(L)$
$L \leftarrow L - \text{Mean}_{DHW}(L)$
$L \leftarrow L/\left(\text{Std}_K(L) + \epsilon\right)$
if training **then**
    $L \leftarrow L + 0.01\,\mathcal{N}(0, I)$
$W \leftarrow \text{Softmax}(L/0.8)$
for $k = 0, \ldots, K-1$ **do**
    $F_k \leftarrow E_k(x)$
    $y \leftarrow \sum_{k=0}^{K-1} W_k \odot F_k$
    $y \leftarrow \text{SiLU}(\text{GroupNorm}(y))$
    return y

Algorithm 2 Decoder Block

Require: Input feature map $x \in \mathbb{R}^{C_{\text{in}}\times D\times H\times W}$
Ensure: Output feature map y

$x \leftarrow \text{Conv}_{1\times1\times1}(x)$
$x_{\text{low}} \leftarrow \text{AvgPool}_{1\times3\times3}(x)$
$x_{\text{high}} \leftarrow x - x_{\text{low}}$
$F_0 \leftarrow \text{Conv}_{1\times1\times1}(x_{\text{low}})$
$F_1 \leftarrow \text{Conv}_{1\times1\times1}(x_{\text{high}})$
$F_2 \leftarrow \text{Conv}_{1\times3\times3}(x)$
$F_3 \leftarrow \text{Conv}_{1\times1\times1}(x)$
$F_4 \leftarrow \text{Conv}_{3\times1\times1}(x)$
$y \leftarrow \frac{1}{5}\sum_{k=0}^{4} F_k$
$y \leftarrow \text{Conv}_{1\times1\times1}(y)$
$y \leftarrow \text{SiLU}(\text{GroupNorm}(y))$
if upsampling is enabled **then**
    $y \leftarrow \text{NearestUpsample}_{H,W}(y, 2)$
return y

Table 3: Mean Reconstruction performance comparison for different models across the test dataset. Lower MSE and higher PSNR and SSIM indicate better performance.

Method	MSE	PSNR (dB)	SSIM
Ours	**0.00147**	**29.04**	**0.955**
MONAI	0.00412	24.05	0.869
ViT	0.00424	24.00	0.877

Table 4: Comparison of model complexity, memory usage, inference time, and high-resolution (HR) reconstruction quality.

Metric	Ours	MONAI	ViT
Parameters	1.293 M	3.003 M	9.925 M
Peak VRAM	2.276 GB	0.313 GB	0.794 GB
Inference Time	347.79 ms	26.35 ms	80.00 ms
PSNR	**29.04 dB**	24.05 dB	24.00 dB

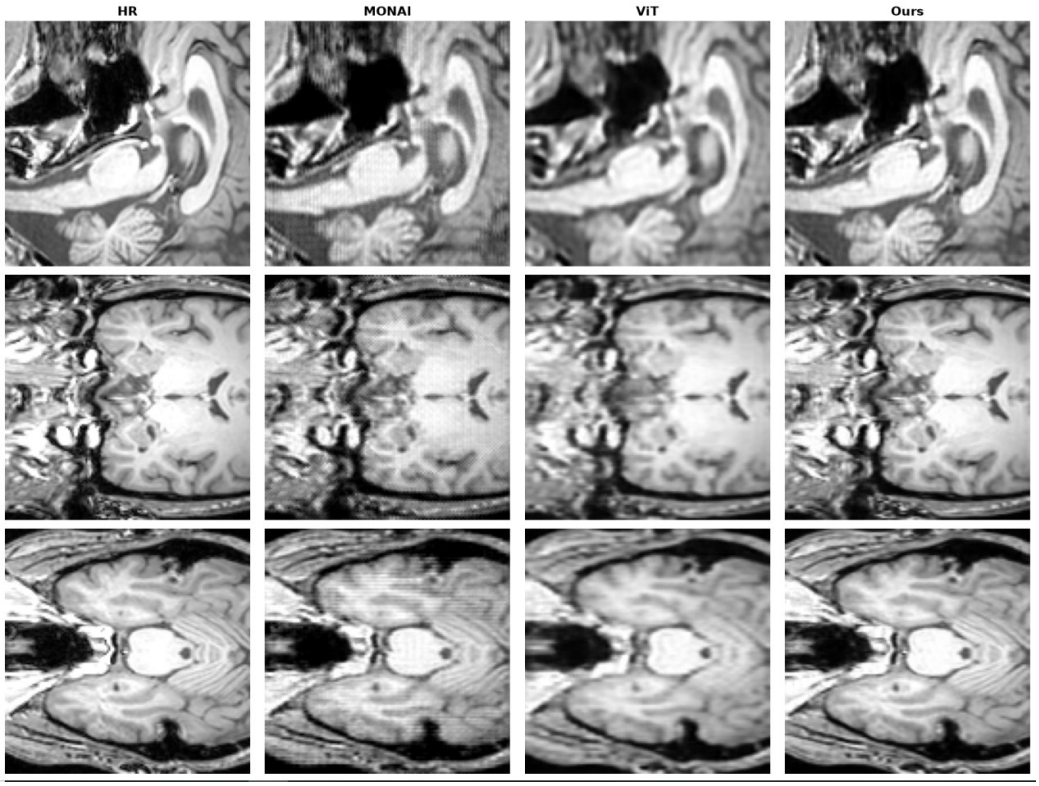


Figure 7: Reconstructions of Different Models and Ground Truth Image

C.2 Qualitative Visualization of Encoder and Decoder Routing

Figure 8 visualizes the spatial routing weights produced by the encoder and decoder routing mechanisms for a representative axial slice. The encoder and decoder produce distinct spatial routing patterns across the five experts, with different experts receiving higher weights in different regions of the feature map.

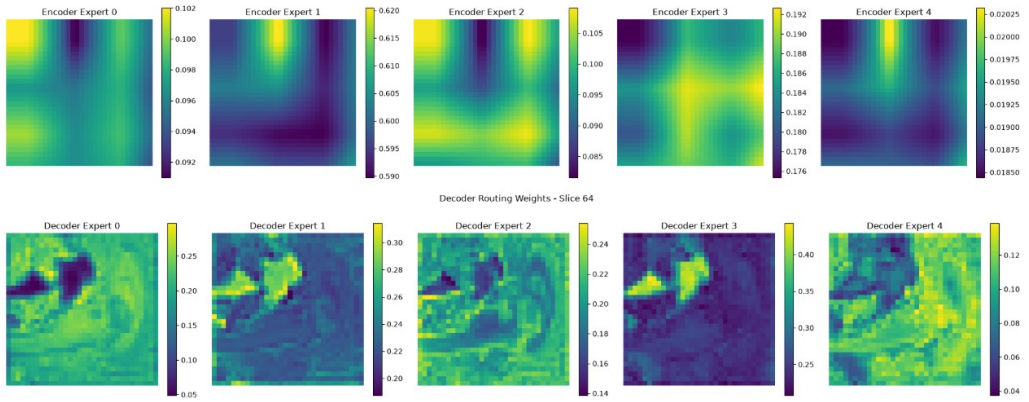


Figure 8: Representative encoder and decoder routing weights for a central axial slice. Each map shows the spatial contribution assigned to one expert. The encoder and decoder exhibit spatially varying routing patterns across the feature map.

The visualizations illustrate that routing weights vary spatially rather than remaining constant across the feature map. Different experts exhibit different spatial patterns, providing a qualitative view of the soft spatial composition analyzed quantitatively in the preceding experiments.

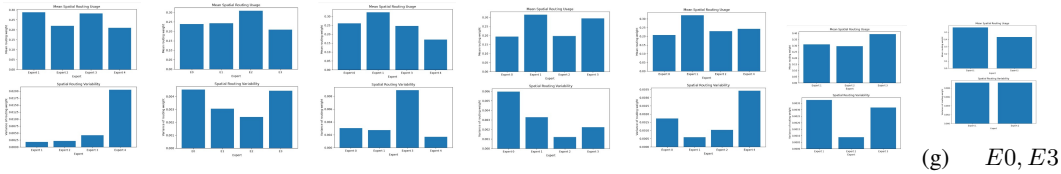
C.3 Reconstruction Ablations

We first evaluate how expert composition and spatial routing affect reconstruction quality. Table 5 reports the reconstruction performance of the full five-expert model, its uniform-routing counterpart, individual expert configurations, and selected reduced expert banks.

Table 5: Mean reconstruction performance across the test set.

Configuration	MSE ↓	PSNR ↑ (dB)	SSIM ↑	Description
Adaptive 5 Experts	0.001310	29.04	0.9596	Full expert bank with spatially adaptive routing
Uniform 5 Experts	0.001368	28.84	0.9586	Full expert bank with uniform expert aggregation
$-E_0$	0.001129	29.68	0.9651	Leave-one-out model without E_0
$-E_1$	0.001561	28.25	0.9534	Leave-one-out model without E_1
$-E_2$	0.003430	24.87	0.8968	Leave-one-out model without E_2
$-E_3$	0.001121	29.64	0.9647	Leave-one-out model without E_3
$-E_4$	0.001528	28.28	0.9551	Leave-one-out model without E_4
$-(E_0, E_3)$	0.001148	29.54	0.9397	Adaptive routing over E_1, E_2 and E_4
$-(E_0, E_3, E_4)$	0.001266	29.14	0.9611	Adaptive routing over E_1 and E_2
E_2	0.001306	29.14	0.9593	Single-expert model using only E_2

The reconstruction results show that performance is not monotonic with the number of experts. The adaptive five-expert model improves over its uniform counterpart, while several reduced expert configurations achieve higher reconstruction quality. In particular, removing E_2 produces a substantial degradation, whereas removing E_0 or E_3 results in similar or higher reconstruction quality than the full five-expert configuration. These results motivate treating expert composition as an impor-



(a) E_0 (b) E_1 (c) E_2 (d) E_3 (e) E_4 (f) $E0$ and E_4 and E_4 (g) $E0, E3$

Figure 9: Mean spatial routing usage and spatial routing variability for the abltations along with the 5 expert model. The experts exhibit non-uniform mean usage and different degrees of spatial variation in their routing weights.

tant factor in the behavior of the model rather than interpreting the number of experts alone as a determinant of reconstruction quality. The reconstruction results are therefore used primarily as supporting evidence, while the remaining analyses focus on the specialization and routing behavior of the experts.

C.4 Spatial Routing Behavior

To further characterize the spatial behavior of the routing mechanism, we measure the mean routing usage and spatial variance of each expert. The mean routing usage measures the average contribution assigned to each expert across the feature map, while the spatial variance measures how strongly that contribution changes across spatial locations.

Figure 9 shows that both expert usage and spatial routing variability depend on the composition of the expert bank. In the full five-expert configuration, routing is non-uniform, with E_3 receiving the largest mean routing weight and E_4 the smallest. Removing individual experts redistributes the routing weights among the remaining operators, and in some configurations produces substantially more concentrated expert usage. The spatial variance of the routing weights likewise differs across experts and changes when the available expert set is modified. Thus, the router does not simply apply a fixed global weighting to the expert bank; instead, the relative contribution and spatial variability of the experts are jointly determined by the learned routing function and the available expert composition.

The unrouted responses exhibit a pronounced correlation pattern for E_1, which shows the strongest positive association across all four structural measures. After routing, this pattern changes substantially. For example, E_0 changes from negative correlations across all four structural measures to positive correlations, while E_4 becomes more negatively associated with the structural measures. The correlations of E_1 remain positive but are reduced for several structural properties.

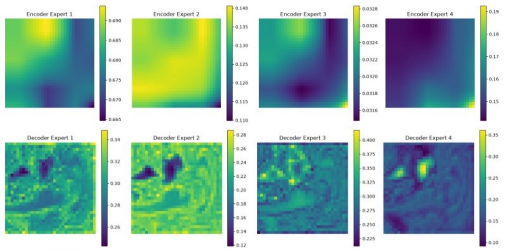
These results indicate that spatial routing changes the relative association of individual expert contributions with the structural characteristics of the input. Importantly, these measurements describe correlation rather than causation: they characterize how the routed contributions are associated with image structure rather than establishing that the router explicitly selects experts based on a particular structural property.

C.5 Structure-Dependent Expert Correlations

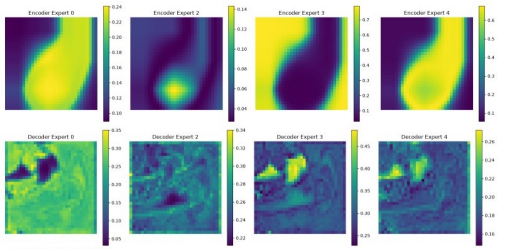
To further examine how expert composition affects the learned specialization patterns, we analyze the relationship between individual expert responses and observable structural properties of the input under different expert ablations. For each configuration, we measure the correlation between the response magnitude of each available expert and four structural quantities computed from the input volume: overall gradient magnitude, in-plane gradient magnitude, through-plane gradient magnitude, and Laplacian magnitude.

For an expert k with feature response F_k, we define its response magnitude at spatial location p as

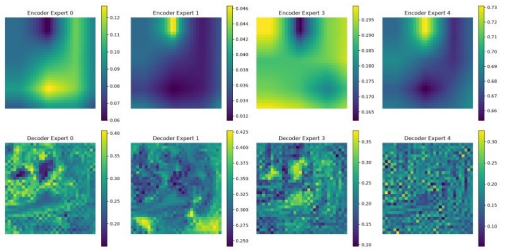
$$A_k(p) = \sqrt{\frac{1}{C}\sum_{c=1}^{C} F_{k,c}(p)^2}. \tag{37}$$



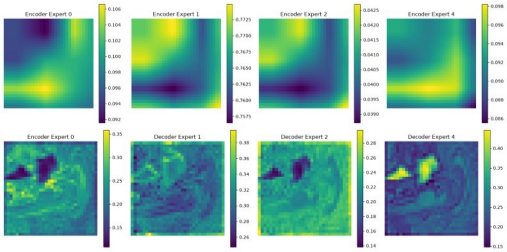
(a) Expert E_0



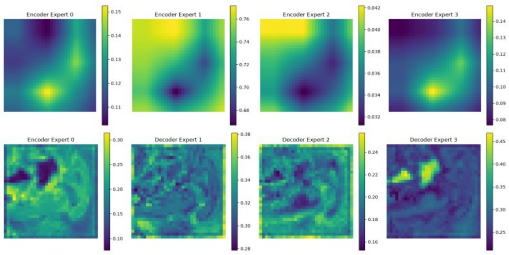
(b) Expert E_1



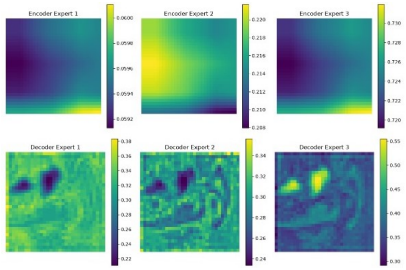
(c) Expert E_2



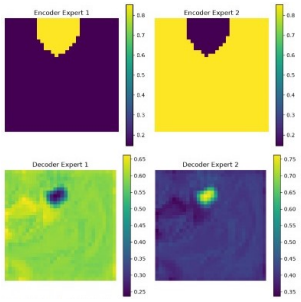
(d) Expert E_3



(e) Expert E_4



(f) Expert $E0$ and E_4



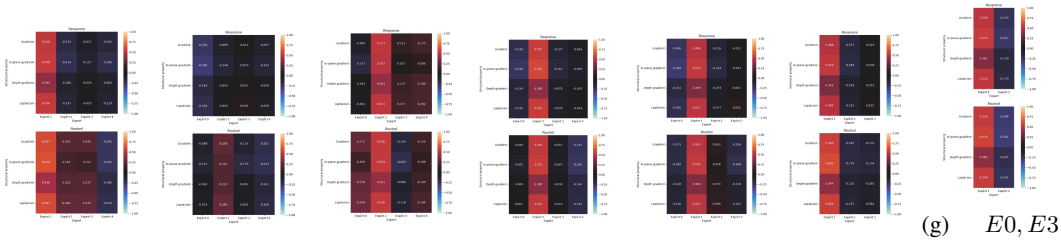
(g) Expert $E0$, $E3$ and E_4

Figure 10: Correlation between expert response magnitude and structural properties before and after spatial routing. Each subfigure shows the corresponding expert. The upper panel reports correlations for the unrouted expert responses, while the lower panel reports correlations for the routed expert contributions.

We then compute the Pearson correlation between A_k and each structural map S over the spatial locations of each test volume and average the resulting correlations across the test set. To distinguish the intrinsic response of an expert from its contribution after routing, we additionally compute the correlation of the routed response

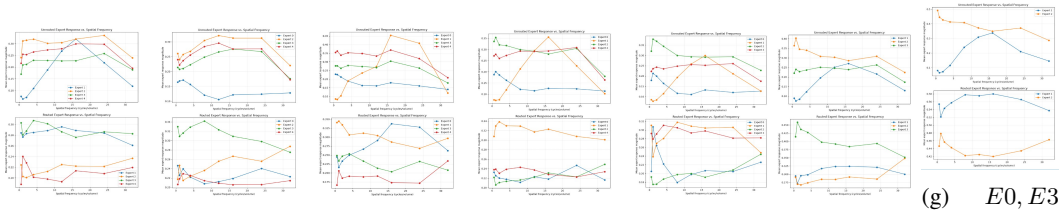
$$C_k(p) = w_k(p)A_k(p), \tag{38}$$

where $w_k(p)$ denotes the spatial routing weight assigned to expert k.



(a) E_0 (b) E_1 (c) E_2 (d) E_3 (e) E_4 (f) $E0$ and E_4 and E_4 (g) $E0, E3$

Figure 11: Structural correlations of the expert responses before and after spatial routing. Each subfigure corresponds to one expert.



(a) E_0 (b) E_1 (c) E_2 (d) E_3 (e) E_4 (f) $E0$ and E_4 and E_4 (g) $E0, E3$

Figure 12: Frequency characteristics of the expert responses before and after spatial routing. Each subfigure corresponds to one expert.

These measurements should be interpreted as associations rather than causal effects. In particular, a positive correlation indicates that an expert's response magnitude tends to be larger in regions with larger values of the corresponding structural quantity; it does not imply that the expert is responsible for generating that structure.

Figure 11 results show that the structural association patterns depend on the available expert composition. In the full expert configuration, the individual experts exhibit substantially different correlation patterns with the structural measures, indicating that the learned operators do not respond uniformly to the same input characteristics. Removing an expert changes both the magnitude and distribution of these correlations among the remaining experts. In several reduced configurations, one remaining expert becomes strongly associated with the gradient- and Laplacian-based measures, while the other experts exhibit weaker or opposing associations.

Importantly, these ablations do not establish that a particular expert is uniquely responsible for a particular structural property. Because the experts are jointly trained and the routing weights form a compositional mixture, removing one operator changes the representation available to the remaining operators and can consequently alter their learned responses. We therefore interpret these experiments as evidence that expert composition influences the resulting specialization, rather than as isolated measurements of individual expert importance.

C.6 Frequency-Dependent Expert Responses

We additionally characterize the frequency-dependent behavior of the experts using controlled sinusoidal inputs. We generate sinusoidal inputs along a fixed in-plane direction while varying the spatial frequency, keeping the input amplitude and spatial orientation fixed. For each frequency, we measure the mean response magnitude of each expert before and after spatial routing.

The unrouted experts exhibit distinct frequency-dependent response profiles. In particular, the experts differ substantially in both their response magnitude and how that magnitude changes across spatial frequencies. Following spatial routing, these response profiles are substantially modified, with the relative contribution of the experts changing across the frequency range.

The results therefore provide a controlled characterization of the frequency-dependent behavior of the heterogeneous expert bank and show that the composition of expert contributions changes when spatial routing is applied. Because the sinusoidal inputs are controlled synthetic stimuli rather than

samples from the training distribution, these results are interpreted as a characterization of the operators and their routing behavior rather than as evidence of a specific frequency-selection mechanism learned from natural images.

C.7 Seed Variability

To assess the sensitivity of the learned expert decomposition to random initialization, we train the five-expert model using three independent random seeds. We observe variation in the resulting spatial routing patterns across runs, indicating that the particular decomposition of functional roles among the expert indices can vary with initialization.

In particular, the spatial routing distributions and the structural associations of individual experts vary across seeds. An expert exhibiting a strong association with a particular structural property in one run need not exhibit the same association in another. We therefore do not interpret individual expert indices as having fixed semantic roles. Instead, our analysis focuses on whether differentiated expert responses and spatially varying expert contributions persist across independent training runs.

The mean absolute response correlation varies from 0.58 to 0.63 across the three seeds, indicating that the overall degree of response redundancy remains comparatively stable despite variation in the identity of individual expert responses.

These results motivate interpreting expert specialization as a property of the learned expert composition rather than as a fixed assignment of semantic functions to individual expert indices. The expert-removal experiments in Section C.5 further examine how this decomposition changes when the available operator set is modified.

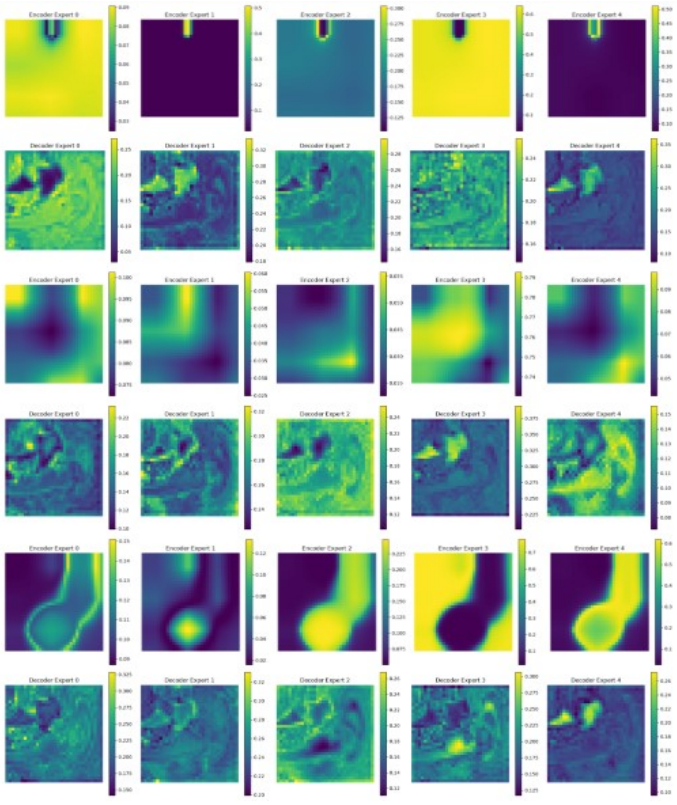


Figure 13: Reconstructions of Different Models and Ground Truth Image

D Experimental Setup and Reproducibility

D.1 Dataset and Preprocessing

We evaluate the proposed formulation on the IXI-T1 dataset using an 80%/10%/10% train-validation-test split at the volume level. All experiments operate on three-dimensional T1-weighted MRI volumes. NIfTI volumes are converted to floating-point arrays and, when necessary, the first volume is selected from four-dimensional NIfTI inputs. The resulting volumes are reordered from $[H, W, D]$ to $[D, H, W]$ before further processing.

To reduce the effect of scanner- and volume-dependent intensity ranges, each volume is independently normalized using percentile-based intensity scaling. Specifically, the 1st and 99th intensity percentiles are used as lower and upper bounds, respectively. Intensities outside this range are clipped, after which the volume is linearly scaled to $[0, 1]$.

For training, volumes are cropped to $128 \times 128 \times 128$ voxels. Random three-dimensional crops are sampled during training, while center crops are used when random cropping is disabled. We additionally apply simple spatial augmentation by independently flipping the volumes along the two in-plane spatial axes with probability 0.5 for each axis.

The standard autoencoder experiments use the normalized high-resolution volume directly as both the input and reconstruction target. Thus, for an input volume x, the training objective is to reconstruct the same $128 \times 128 \times 128$ volume.

D.2 Training Configuration

All autoencoder variants are trained using the same reconstruction objective and optimization procedure. Models are trained for 20 epochs using the same training and validation splits. The five-expert configuration is used as the reference model, while reduced expert sets are used for the expert-composition ablations.

Models are optimized using AdamW with a learning rate of 1×10^{-4}. No learning-rate scheduler is used. Given an input volume x and its reconstruction $\hat{x}$, the training objective is the voxel-wise mean squared error

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - x_i)^2, \tag{39}$$

where N is the number of voxels in the input volume. The same reconstruction objective is used during validation.

Models are trained from random initialization. For experiments involving multiple random seeds, a separate model instance is initialized for each seed. The seed-variability analysis uses three independent random seeds, as described in Appendix C.7.

D.3 Evaluation Protocol

All reconstruction metrics are computed on the held-out test volumes without masking. For each test volume, the model produces a reconstructed volume $\hat{x}$ from the input volume x. We compute MSE, PSNR, and three-dimensional SSIM independently for each volume and subsequently aggregate the resulting values across the test set.

The mean squared error for a test volume is computed as

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - x_i)^2, \tag{40}$$

where N is the number of voxels in the volume. PSNR is computed from the per-volume MSE using a data range of 1:

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{1}{\mathrm{MSE}} \right). \tag{41}$$

For each volume, we additionally compute three-dimensional structural similarity (SSIM) using the same data range. The final reported values are the mean and standard deviation across test volumes. Median values are also retained for additional analysis.

All three metrics are computed independently for each volume before aggregation, such that each test volume contributes equally to the reported statistics. No test-time masking or voxel selection is applied.

In addition to reconstruction quality, we evaluate the behavior of the individual convolutional experts and the spatial routing mechanism. For an expert k, its response magnitude at spatial location p is defined as

$$A_k(p) = \sqrt{\frac{1}{C}\sum_{c=1}^{C} F_{k,c}(p)^2}, \tag{42}$$

where F_k denotes the feature response produced by expert k and C is the number of output channels. The corresponding routed contribution magnitude is

$$C_k(p) = w_k(p)A_k(p), \tag{43}$$

where $w_k(p)$ is the spatial routing weight assigned to expert k.

For expert-response redundancy, we compute the Pearson correlation between pairs of expert response maps and report the mean absolute pairwise correlation,

$$R = \frac{2}{K(K-1)}\sum_{i<j} |\rho_{ij}|\,, \tag{44}$$

where K is the number of available experts and ρ_{ij} denotes the Pearson correlation between the response maps of experts i and j. Higher values indicate greater similarity between expert responses.

We additionally measure the entropy of the routing distribution to characterize the concentration of the expert mixture. Given the routing weights $\mathbf{w}(p)$, the normalized entropy is

$$H(p) = -\frac{1}{\log K}\sum_{k=1}^{K} w_k(p)\log w_k(p). \tag{45}$$

The entropy is averaged over spatial locations and test volumes. Higher values correspond to a more distributed routing mixture, whereas lower values indicate a more concentrated distribution.

For structural analyses, we compute gradient magnitude, in-plane gradient magnitude, through-plane gradient magnitude, and Laplacian magnitude directly from the input volume. Pearson correlations between these structural maps and both $A_k(p)$ and $C_k(p)$ are computed over spatial locations and averaged across the test set. These measurements are interpreted as spatial associations rather than causal relationships.

Finally, we characterize the frequency response of the individual experts using controlled three-dimensional sinusoidal inputs with varying spatial frequency. For a frequency f, the input is defined as

$$x_f(x,y,z) = A\sin\left(\frac{2\pi f x}{W} + \phi\right), \tag{46}$$

where W is the volume width and A and ϕ denote the amplitude and phase. We measure both the response magnitude of each expert and the corresponding mean routing weights across the volume. This experiment provides a controlled characterization of frequency-dependent expert behavior rather than an evaluation on natural MRI inputs.