Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

Structured Modality-Aware Token Interaction for Multimodal Medical Imaging

Selene Tomassini[1*], **Hafiza Ayesha Hoor Chaudhry**[1], **Alessandro Galdelli**[2] and **Paolo Giorgini**[1]

[1]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[2]Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy
{selene.tomassini, hafiza.chaudhry}@unitn.it, a.galdelli@univpm.it, paolo.giorgini@unitn.it

Abstract

Multimodal medical imaging benefits from the global context modeling of transformers, yet most existing models fuse modalities implicitly by channel concatenation, leaving cross-modal interaction unstructured or relying on costly multi-stream cross-attention. We propose Modality-Aware Token Interaction (MATI), an architecturally lightweight and backbone-agnostic module that structures multimodal interaction within a single token stream by partitioning embedding channels into modality-aligned subspaces. MATI performs modality-preserving intra-subspace self-attention and gated global mixing for controlled inter-subspace exchange. We instantiate MATI in UNETR and introduce two architectures: ModaUNETR^S refines selected skip-token representations, and ModaUNETR^E injects MATI into the transformer encoder to progressively shape the token hierarchy. Experiments on the BraTS 2020 benchmark employ five-fold cross-validation and report segmentation and efficiency metrics on the official training and validation sets. Both models consistently improve over UNETR across all tumor subregions, with larger gains for tumor core and whole tumor, and ModaUNETR^E further improves upon ModaUNETR^S. An ablation study confirms monotonic gains from structured interaction. Compared with heavier transformers, the proposed models achieve a favorable accuracy-efficiency trade-off without modality-specific encoders or quadratic cross-attention, supporting structured multimodal interaction as a first-class architectural principle. Implementation of MATI and its instantiations is available at https://github.com/S3l11/MATI.

1 Introduction

Transformer architectures have reshaped representation learning by enabling explicit modeling of long-range dependencies through self-attention [Vaswani *et al.*, 2017]. In medical imaging, this property is particularly valuable for volumetric data, where global anatomical context plays a critical role. When applied to multimodal inputs, however, most existing architectures still rely on implicit fusion, typically implemented by concatenating modalities along the channel dimension at the network input. This strategy assumes that heterogeneous modalities can be processed within a single shared representation space without explicit structural constraints. In practice, standard self-attention tends to conflate modality-specific feature learning and cross-modal fusion into a single unstructured operation, forcing the network to simultaneously disentangle modality-specific representations and learn how to combine them. As a result, multimodal interaction remains implicit and difficult to control.

Several recent works have explored explicit modality-aware designs based on multi-branch encoders, cross-attention mechanisms, or modality-specific transformer streams [Xing *et al.*, 2022; Zhang *et al.*, 2022; Sun, 2025]. Although effective, these approaches typically increase architectural complexity, memory consumption, and computational cost, and frequently rely on quadratic cross-attention between modality-specific token sets. This makes it difficult to disentangle the benefits of architectural inductive bias from those of increased model capacity. These limitations are particularly evident in multimodal 3D medical image segmentation, where accurate delineation requires the coherent integration of complementary sequences over large spatial contexts. In this setting, transformer-based models such as UNETR [Hatamizadeh *et al.*, 2022b] and SwinUNETR [Hatamizadeh *et al.*, 2022a] have demonstrated strong performance, but still largely rely on implicit or heavyweight fusion mechanisms.

In this work, we argue that structured modality interaction should be treated as a first-class architectural design principle in multimodal learning. We propose Modality-Aware Token Interaction (MATI), an architecturally lightweight and backbone-agnostic module that operates directly on token embeddings by structuring the channel space into modality-aligned subspaces. MATI explicitly separates modality-preserving refinement from controlled cross-modal information exchange, injecting a structured multimodal inductive bias into a single-stream transformer representation. We instantiate MATI within the UNETR framework and propose two architectures: ModaUNETR^S and ModaUNETR^E. ModaUNETR^S applies MATI to selected intermediate to-

*Corresponding author

ken representations used for skip connections, whereas ModaUNETR^E injects MATI directly inside the transformer encoder to progressively shape the representation hierarchy. Both variants preserve the overall UNETR topology and introduce only a moderate computational overhead. We evaluate the proposed models on the BraTS 2020 benchmark, a challenging and well-established testbed for multimodal representation learning in medical imaging.

In summary, our contributions are threefold:

- We introduce MATI, an architecturally lightweight and backbone-agnostic modality-aware interaction module that structures multimodal fusion inside the token embedding space.
- We propose ModaUNETR^S and ModaUNETR^E, two instantiations of MATI within a UNETR backbone using respectively skip-level and in-encoder interaction.
- We demonstrate on BraTS 2020 that explicitly structuring multimodal interaction improves accuracy without substantially increasing computational complexity or introducing additional attention streams.

2 Related Work

Convolutional encoder-decoder architectures such as U-Net and its 3D variants [Ronneberger *et al.*, 2015; Çiçek *et al.*, 2016] have long dominated medical image segmentation. Frameworks such as nnU-Net [Isensee *et al.*, 2021] further showed that carefully engineered pipelines can achieve strong and reproducible performance. However, convolution-based models remain inherently limited by local receptive fields, reducing their ability to capture long-range spatial dependencies and global anatomical context.

Transformer-based architectures address this limitation through self-attention mechanisms that explicitly model long-range interactions. Following the success of Vision Transformers (ViTs) [Dosovitskiy *et al.*, 2020], several transformer-based models have been proposed for medical image segmentation, including UNETR [Hatamizadeh *et al.*, 2022b], TransUNet [Chen *et al.*, 2021], and Swin-UNETR [Hatamizadeh *et al.*, 2022a]. These works established transformers as a powerful alternative to purely convolutional designs for volumetric segmentation.

In multimodal medical image segmentation, however, most transformer-based approaches still rely on implicit fusion by concatenating modalities at the input level [Myronenko, 2019; Fareed *et al.*, 2025]. While effective, this strategy treats modalities as homogeneous information sources and offers no explicit control over modality-specific representation and interaction. To overcome this limitation, several works introduced explicit modality-aware designs based on multi-branch encoders, cross-attention, or auxiliary objectives, including mmFormer [Zhang *et al.*, 2022], NestedFormer [Xing *et al.*, 2022], and cross-modality distillation approaches [Lin *et al.*, 2023]. Although these methods often improve performance, they typically require multiple modality-specific streams or full cross-attention modules, substantially increasing architectural complexity and computational cost.

In contrast, our approach embeds modality-aware structure directly within a single-stream transformer by factorizing the token embedding space into modality-aligned subspaces and regulating cross-modal interaction within it. This preserves the simplicity and scalability of architectures such as UNETR while introducing an explicit multimodal inductive bias, offering a lightweight alternative to both implicit early fusion and complex multi-stream transformer designs.

3 Methodology

3.1 Backbone

ModaUNETR^S and ModaUNETR^E are built upon UNETR [Hatamizadeh *et al.*, 2022b], a transformer-based encoder-decoder architecture for volumetric medical image segmentation that combines global context modeling through a ViT encoder with high-resolution reconstruction via a convolutional decoder.

Following the standard UNETR configuration, input volumes of size $128 \times 128 \times 128$ are partitioned into non-overlapping $16 \times 16 \times 16$ patches, producing 512 tokens embedded in a 768-dimensional space and processed by a 12-layer transformer encoder with 12 attention heads and MLP dimension 3072. Intermediate token representations are projected into spatial feature maps and forwarded to a hierarchical convolutional decoder through skip connections for multi-scale feature fusion. This design contains approximately 1.02×10^8 parameters and is kept unchanged across all experiments, providing a stable backbone for isolating the effect of the proposed modality-aware interaction.

3.2 ModaUNETR^S

In multimodal medical imaging, UNETR processes different modalities by concatenating them along the channel dimension before tokenization. After patch embedding, modality information becomes fully entangled within the token embeddings, and all subsequent transformer layers operate in a shared representation space. Although effective, this design provides no explicit mechanism to preserve modality-specific structure or regulate cross-modal information exchange.

ModaUNETR^S introduces an explicit yet low-overhead modality-structured inductive bias by augmenting UNETR with MATI blocks applied to selected transformer skip representations. MATI does not modify the ViT encoder itself. Instead, it refines intermediate token representations before they are projected and injected into the convolutional decoder through the standard UNETR skip connections.

Placement. As illustrated in Figure 1, MATI blocks are applied to the transformer outputs Z_6 and Z_9, corresponding to the outputs of the 6^{th} and 9^{th} ViT encoder blocks. These token sequences are the same representations used by UNETR to build two of the multi-scale skip connections into the convolutional decoder, hence $\mathcal{I}_{\text{MATI}} = \{6, 9\}$. This placement targets intermediate semantic representations, where global contextual information has already been established but before token-to-feature-map projection and spatial decoding. At these depths, tokens capture high-level semantic information while still preserving localized spatial correspondence, making them suitable for modality-aware refinement without interfering with either early low-level processing or late-stage spatial reconstruction.

Subspace-structured interaction. Let $Z \in \mathbb{R}^{B\times N\times C}$ denote a transformer token sequence selected for refinement. MATI introduces an explicit modality-structured inductive bias by factorizing the channel dimension into M fixed subspaces, each associated with one input modality. In practice, we use $M = 4$ subspaces corresponding to the four Magnetic Resonance Imaging (MRI) modalities:

$$Z = \mathrm{Concat}\big(Z^{(1)}, \ldots, Z^{(M)}\big).$$

MATI then performs two complementary operations: intra-subspace refinement, which preserves modality-aligned processing through independent self-attention within each subspace, and inter-subspace mixing, which enables controlled cross-modal information exchange via a lightweight global gating mechanism. The complete MATI operation is summarized in Algorithm 1.

Intra- vs. inter-subspace. For each modality subspace, MATI applies a residual self-attention block:

$$U^{(m)} = Z^{(m)} + \mathrm{MHA}_m(\mathrm{LN}(Z^{(m)})), \quad m = 1, \ldots, M.$$

This allows token interaction within each modality-aligned subspace while preserving structured multimodal representations.

To enable controlled cross-modality interaction without introducing full cross-attention over all tokens, MATI summarizes each refined subspace by average pooling over the token dimension, concatenates the resulting summary vectors, and predicts a set of mixing weights through a small gating network. These weights are used to form a global mixed feature that is broadcast back to all subspaces and added residually to each token representation. The final refined token sequence is then reconstructed by concatenation:

$$Z' = \mathrm{Concat}\big(\hat{Z}^{(1)}, \ldots, \hat{Z}^{(M)}\big).$$

This design explicitly factorizes modality-preserving refinement from controlled modality interaction, instantiating a structured multimodal inductive bias while operating entirely within the original UNETR token space.

Complexity and integration. MATI introduces no additional token streams, parallel modality-specific encoders, or quadratic cross-attention across modalities. It preserves the token sequence length N and is applied only to a small number of skip representations. The additional parameters are limited to the per-subspace attention projections and the lightweight mixing network. Consequently, performance gains can be attributed to the introduced structural bias rather than a substantial increase in backbone capacity, whereas computational complexity remains dominated by the original UNETR architecture. After MATI refinement, the updated token representations are projected into spatial feature maps through the standard UNETR token-to-feature projection layers and injected into the convolutional decoder via unchanged skip connections. The complete forward pass of ModaUNETR^S is summarized in Algorithm 2.

3.3 ModaUNETR^E

While ModaUNETR^S applies MATI only to selected transformer outputs later used as UNETR skip tokens,

Algorithm 1 Modality-Aware Token Interaction (MATI).

Require: Token sequence $Z \in \mathbb{R}^{B\times N\times C}$, number of modalities M
Ensure: Refined token sequence $Z' \in \mathbb{R}^{B\times N\times C}$
// Split channels into modality-aligned subspaces
$\{Z^{(1)}, \ldots, Z^{(M)}\} \leftarrow \mathrm{SplitChannels}(Z)$
// Intra-subspace refinement
for $m = 1$ to M **do**
  $U^{(m)} \leftarrow Z^{(m)} + \mathrm{MHA}_m(\mathrm{LN}(Z^{(m)}))$
end for
// Compute subspace summaries
for $m = 1$ to M **do**
  $s^{(m)} \leftarrow \mathrm{Mean}_n(U^{(m)})$
end for
$s \leftarrow \mathrm{Concat}(s^{(1)}, \ldots, s^{(M)})$
// Predict mixing weights
$\alpha \leftarrow \sigma(\mathrm{MLP}(s))$
// Inter-subspace mixing
$G \leftarrow \sum_{m=1}^{M} \alpha_m \cdot U^{(m)}$ {broadcasted to all tokens}
// Residual fusion and reconstruction
for $m = 1$ to M **do**
  $\hat{Z}^{(m)} \leftarrow U^{(m)} + G$
end for
$Z' \leftarrow \mathrm{Concat}(\hat{Z}^{(1)}, \ldots, \hat{Z}^{(M)})$
return Z'

Algorithm 2 ModaUNETR^S forward pass with MATI skip-token refinement.

Require: Multimodal input $X \in \mathbb{R}^{B\times M\times D\times H\times W}$
Ensure: Segmentation prediction $\hat{Y}$
// ViT encoder (standard UNETR)
$\{Z_1, \ldots, Z_{12}\}, Z_{\text{out}} \leftarrow \mathrm{ViT}(X)$
// Extract UNETR skip tokens
$S \leftarrow \{Z_3, Z_6, Z_9, Z_{12}\}$
// MATI refinement on selected skip tokens
// MATI refinement indices: $\mathcal{I}_{\text{MATI}} = \{6, 9\}$
for $\ell \in \mathcal{I}_{\text{MATI}}$ **do**
  $Z_\ell \leftarrow \mathrm{MATI}(Z_\ell)$
end for
// Token-to-feature projection (standard UNETR)
$E_1 \leftarrow \mathrm{Encoder}_1(X)$
$E_2 \leftarrow \mathrm{Encoder}_2(\mathrm{Proj}(Z_3))$
$E_3 \leftarrow \mathrm{Encoder}_3(\mathrm{Proj}(Z_6))$ {Z_6 is MATI-refined}
$E_4 \leftarrow \mathrm{Encoder}_4(\mathrm{Proj}(Z_9))$ {Z_9 is MATI-refined}
$E_5 \leftarrow \mathrm{Proj}(Z_{\text{out}})$
// Convolutional decoder (standard UNETR)
$D_4 \leftarrow \mathrm{Decoder}_5(E_5, E_4)$
$D_3 \leftarrow \mathrm{Decoder}_4(D_4, E_3)$
$D_2 \leftarrow \mathrm{Decoder}_3(D_3, E_2)$
$D_1 \leftarrow \mathrm{Decoder}_2(D_2, E_1)$
$\hat{Y} \leftarrow \mathrm{Out}(D_1)$
return $\hat{Y}$

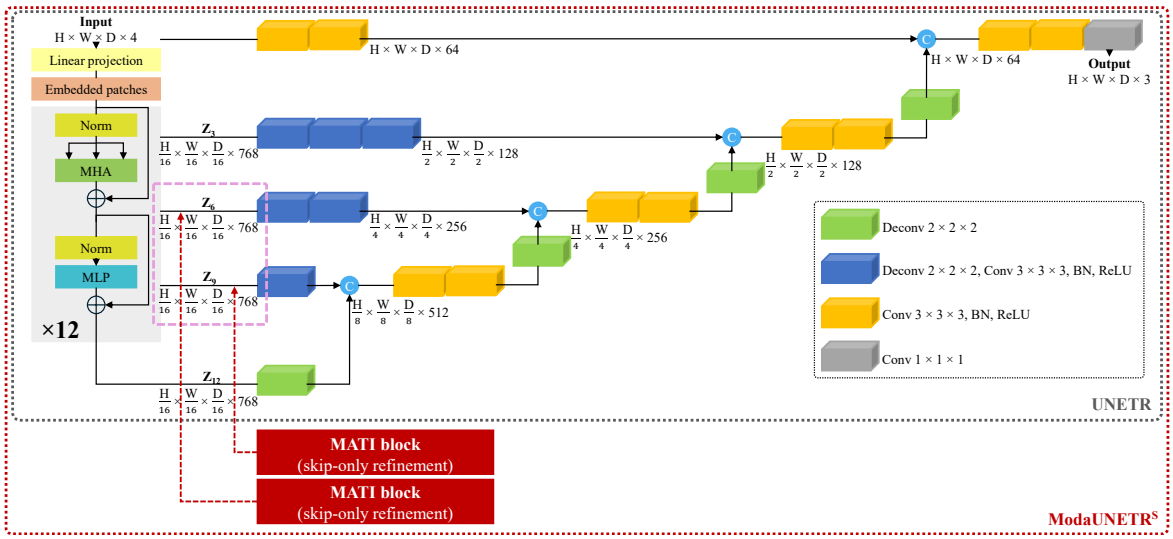

Figure 1: Overview of ModaUNETR^S. The UNETR backbone is preserved, while MATI blocks refine the intermediate token representations Z_6 and Z_9 used for skip connections before token-to-feature projection. The ViT encoder, skip topology, and convolutional decoder remain unchanged; only the selected skip tokens are refined via subspace-structured interaction.

ModaUNETR^E injects the same interaction mechanism directly inside the ViT encoder, allowing modality-aware interaction to shape the evolving token stream.

In-encoder MATI injection. Let $\{B_0, \ldots, B_{11}\}$ denote the twelve transformer blocks of the ViT encoder and let $Z_0 = \mathrm{PatchEmbed}(X)$ be the embedded token sequence. ModaUNETR^E modifies the ViT forward pass by inserting MATI after the 3^{rd} and 6^{th} blocks, as illustrated in Figure 2. In our implementation, this corresponds to the 0-indexed set $\mathcal{J}_{\text{MATI}} = \{2, 5\}$. The token stream update becomes:

$$Z_{i+1} = \begin{cases} \mathrm{MATI}(B_i(Z_i)) & \text{if } i \in \mathcal{J}_{\text{MATI}}, \\ B_i(Z_i) & \text{otherwise}, \end{cases} \qquad i = 0, \ldots, 11.$$

When applied, MATI updates the running token stream, which is then propagated through all subsequent transformer blocks. As a result, downstream layers operate on modality-structured tokens rather than on the original entangled representation.

Skip extraction from modified stream. UNETR uses four transformer outputs as skip tokens, then projected to feature maps. Consistent with the UNETR design, we tap the token stream at block indices $\{2, 5, 8, 11\}$, yielding $\{Z_3, Z_6, Z_9, Z_{12}\}$ from the MATI-refined stream. These representations are projected into spatial feature maps and injected into the unchanged convolutional decoder through the standard UNETR skip connections.

Complexity and integration. ModaUNETR^E preserves the overall UNETR topology (encoder-decoder structure, skip scheme, and token-to-feature projections) and only augments the ViT forward pass with two MATI modules. Consequently, the parameter and computational overhead remain limited, while modality-aware interaction influences both the decoder inputs and the internal transformer representation hierarchy. The complete forward pass of ModaUNETR^E is summarized in Algorithm 3.

Algorithm 3 ModaUNETR^E forward pass with MATI injected inside the ViT encoder.

Require: Multimodal input $X \in \mathbb{R}^{B \times M \times D \times H \times W}$
Ensure: Segmentation prediction $\hat{Y}$
// Patch embedding (standard UNETR ViT front-end)
$Z \leftarrow \mathrm{PatchEmbed}(X)$
$H \leftarrow [\,]$ {list of tapped skip tokens}
// ViT encoder blocks with MATI injection
for $i = 0$ to 11 **do**
  $Z \leftarrow B_i(Z)$ {standard transformer block}
  if $i \in \mathcal{J}_{\text{MATI}}$ **then**
    $Z \leftarrow \mathrm{MATI}(Z)$ {$\mathcal{J}_{\text{MATI}} = \{2, 5\}$ (after 3^{rd} and 6^{th} blocks)}
  end if
  if $i \in \{2, 5, 8, 11\}$ **then**
    append Z to H {$H = [Z_3, Z_6, Z_9, Z_{12}]$ from the *modified* token stream}
  end if
end for
$Z_{\text{out}} \leftarrow \mathrm{LN}(Z)$
// Token-to-feature projection (standard UNETR)
$E_1 \leftarrow \mathrm{Encoder}_1(X)$
$E_2 \leftarrow \mathrm{Encoder}_2(\mathrm{Proj}(H[0]))$ {Z_3}
$E_3 \leftarrow \mathrm{Encoder}_3(\mathrm{Proj}(H[1]))$ {Z_6}
$E_4 \leftarrow \mathrm{Encoder}_4(\mathrm{Proj}(H[2]))$ {Z_9}
$E_5 \leftarrow \mathrm{Proj}(Z_{\text{out}})$
// Convolutional decoder (standard UNETR)
$D_4 \leftarrow \mathrm{Decoder}_5(E_5, E_4)$
$D_3 \leftarrow \mathrm{Decoder}_4(D_4, E_3)$
$D_2 \leftarrow \mathrm{Decoder}_3(D_3, E_2)$
$D_1 \leftarrow \mathrm{Decoder}_2(D_2, E_1)$
$\hat{Y} \leftarrow \mathrm{Out}(D_1)$
return $\hat{Y}$

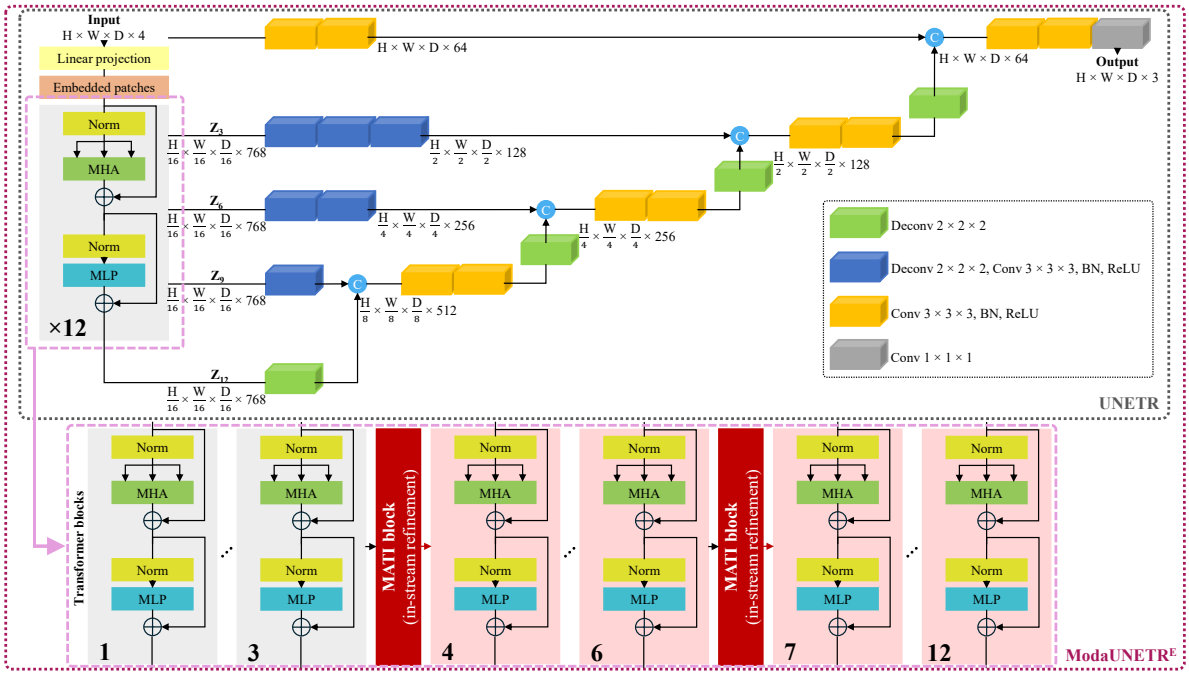

Figure 2: Overview of ModaUNETR^E. MATI is injected into the ViT encoder after the 3^{rd} and 6^{th} transformer blocks ($i = \{2, 5\}$), and the refined tokens propagate through subsequent layers. Skip representations $\{Z_3, Z_6, Z_9, Z_{12}\}$ are extracted from the modified token stream and decoded using the unchanged UNETR projection and convolutional decoder.

4 Experiments

4.1 Data

All experiments are conducted on the BraTS 2020 benchmark [Menze *et al.*, 2015; Bakas *et al.*, 2017; Spyridon *et al.*, 2017; Bakas *et al.*, 2018], which provides pre-operative multimodal MRI scans of gliomas acquired across multiple institutions and imaging protocols. Each case includes four MRI sequences: T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2), and FLAIR. Ground-truth annotations define three tumor subregions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT).

We use the official training set with five-fold cross-validation and patient-level splits, and report state-of-the-art comparisons on the BraTS 2020 validation set.

4.2 Preprocessing

Volumes are provided co-registered, resampled to 1 mm^3 isotropic resolution, and skull-stripped. For computational efficiency and architectural compatibility, all inputs are further resized to $128 \times 128 \times 128$. Image intensities are resampled using trilinear interpolation, while segmentation labels use nearest-neighbor interpolation to preserve discrete class values.

4.3 Training Protocol

All models are implemented in PyTorch and trained on an NVIDIA A100 with mixed precision under a unified setup.

We use Adam with initial learning rate 10^{-4}, decayed by a factor of 0.5 every 10 epochs, and train for 100 epochs per fold with batch size 1. Standard 3D augmentations (flips, affine transforms, intensity perturbations, and bias-field augmentation) are applied only to training volumes. The loss combines voxel-wise cross-entropy (0.3) and soft Dice loss (0.7).

4.4 Evaluation Protocol

Performance is evaluated using Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and Volume Similarity (VS) under five-fold cross-validation on the BraTS 2020 training set.

For state-of-the-art comparison, we report DSC on the BraTS 2020 validation set following standard benchmark practice. Computational efficiency is evaluated on the same validation set in terms of peak GPU memory consumption and inference time, providing a practical measure of the accuracy-cost trade-off across architectures.

4.5 Baseline Comparison

We compare ModaUNETR^S and ModaUNETR^E against representative convolutional and transformer-based architectures for 3D medical image segmentation. UNETR [Hatamizadeh *et al.*, 2022b] serves as the primary transformer baseline and backbone reference. SwinUNETR [Hatamizadeh *et al.*, 2022a] is included as a hierarchical transformer with shifted-window attention, representing a strong high-capacity design, whereas UNet++ [Zhou *et al.*, 2018] provides a competitive convolutional baseline with nested skip connections.

All models are trained and evaluated under the same preprocessing, training, and evaluation protocols to ensure fair comparison.

4.6 Results

Quantitative vs. qualitative. Table 1 reports DSC, IoU, and VS for all subregions under five-fold cross-validation on the BraTS 2020 training set, together with DSC and efficiency metrics (GPU memory and inference time) on the official, held-out validation set. Figure 3 shows representative ground truth and predicted segmentations on the BraTS 2020 training set under five-fold cross-validation.

SwinUNETR performs very well on all subregions, but at higher computational cost. UNet++ yields the lowest resource usage but also the weakest segmentation performance, whereas UNETR provides a strong intermediate baseline. Both ModaUNETR^S and ModaUNETR^E consistently outperform UNETR across all subregions, with larger gains on TC and WT. In particular, ModaUNETR^E further improves over ModaUNETR^S while maintaining substantially lower inference time and memory footprint than SwinUNETR. The qualitative results mirror the quantitative findings. SwinUNETR produces spatially coherent segmentations but oversmooths boundaries. UNet++ often misses small or disconnected enhancing areas, whereas UNETR shows occasional under-segmentation. In contrast, ModaUNETR^S yields more consistent tumor delineations with reduced fragmentation, and ModaUNETR^E further improves structural coherence and tumor extent coverage in challenging cases.

State-of-the-art comparison. Table 2 directly compares ModaUNETR^S and ModaUNETR^E against representative state-of-the-art methods on the BraTS 2020 validation set in terms of DSC for ET, TC, and WT. One compared method [Wang *et al.*, 2020] corresponds to a top-performing BraTS 2020 challenge entry, while the others [An *et al.*, 2024; Kunjumon *et al.*, 2025; Zhou *et al.*, 2025] represent recent architectural trends.

Both ModaUNETR^S and ModaUNETR^E achieve competitive performance across all subregions. In particular, ModaUNETR^E attains the best or tied-best DSC on ET and TC and matches the top WT performance, while ModaUNETR^S remains consistently close to the strongest competing methods.

4.7 Ablation Study

To isolate the effect of architectural structure from parameter count, we compare UNETR, ModaUNETR^S, and ModaUNETR^E under controlled settings. UNETR serves as the baseline without explicit modality-aware interaction. ModaUNETR^S applies MATI to selected skip-token representations, whereas ModaUNETR^E injects MATI inside the ViT encoder for progressive token refinement. All models share the same backbone and differ only in the presence and placement of MATI.

As shown in Table 3, ModaUNETR^S consistently improves over UNETR across all subregions, while ModaUNETR^E further improves over ModaUNETR^S with only moderate increases in memory usage and inference time. Since no modality-specific encoders, additional token streams, or cross-attention mechanisms are introduced, these gains cannot be attributed to increased capacity alone. Instead, they suggest that progressively deeper modality-aware interaction improves performance, with in-encoder injection proving even more effective than skip-level refinement.

5 Discussion

This work investigates whether multimodal interaction in medical imaging can be improved through an explicit architectural inductive bias, without relying on multi-branch encoders or quadratic cross-attention between modality-specific token streams.

Interaction inside token embedding space. MATI structures multimodal interaction directly within the token embedding space by factorizing channels into modality-aligned subspaces. This explicitly separates two roles often entangled in standard self-attention: intra-subspace attention preserves coherent modality-specific representations instead of immediately merging them into a shared space, while inter-subspace mixing enables controlled global interaction via a compact gating mechanism. This separation reduces the burden on generic self-attention to jointly infer modality structure and semantic fusion.

Skip-level vs. in-encoder injection. In ModaUNETR^S, MATI refines only selected transformer outputs used as skip tokens, leaving the ViT encoder unchanged. In contrast, ModaUNETR^E injects MATI directly into the ViT encoder, allowing the refined token stream to propagate through subsequent transformer blocks and influence the entire representation hierarchy. The consistent improvement of ModaUNETR^E over ModaUNETR^S across all folds and subregions (Table 1) suggests that early and intermediate structured interaction is more effective than applying modality-aware refinement only at the decoder interface.

Accuracy-efficiency trade-off. MATI-equipped models provide a favorable middle ground, narrowing the accuracy gap with high-capacity transformers such as SwinUNETR while maintaining lower resource usage (Table 1). This trend is further supported by the state-of-the-art comparison (Table 2), where ModaUNETR^E matches or achieves top performance despite using a simpler single-stream architecture. These results indicate that structured modality-aware interaction can recover much of the performance of heavier multimodal designs at lower computational burden.

Structure vs. capacity. A central motivation of this work is to disentangle architectural inductive bias from brute-force scaling. The ablation study (Table 3) comparing UNETR, ModaUNETR^S, and ModaUNETR^E shows monotonic performance gains despite only moderate increases in memory usage and inference time. Since these improvements are achieved without increasing the number of tokens, introducing additional attention streams, or using modality-specific encoders, the gains are likely driven by the imposed architectural structure rather than increased model capacity.

Stability and robustness. Although we do not conduct a dedicated multi-seed variance analysis, the improvements of ModaUNETR^S and ModaUNETR^E over UNETR are consistent across all folds and subregions (Tables 1 and 3), and we did not observe significant fold-to-fold variance or unstable training despite batch size 1.

Model	Five-Fold Cross-Validation									Held-Out Validation								
	ET			TC			WT			ET			TC			WT		
	DSC↑	IoU↑	VS↑	DSC↑	IoU↑	VS↑	DSC↑	IoU↑	VS↑	DSC↑	GB	s	DSC↑	GB	s	DSC↑	GB	s
UNet++	0.70±.03	0.65±.04	0.81±.03	0.83±.01	0.77±.03	0.92±.02	0.76±.02	0.69±.03	0.84±.01	0.70	1.54	0.05	0.82	1.54	0.05	0.75	1.54	0.05
UNETR	0.73±.01	0.67±.02	0.85±.02	0.85±.02	0.81±.02	0.94±.01	0.78±.02	0.72±.01	0.88±.01	0.71	0.36	0.03	0.84	0.36	0.03	0.78	0.36	0.03
SwinUNETR	0.81±.04	0.74±.03	0.92±.01	**0.91**±.01	0.85±.02	**0.97**±.02	0.87±.03	**0.81**±.03	0.94±.02	**0.81**	4.05	0.14	0.90	4.05	0.14	**0.87**	4.05	0.14
ModaUNETR^S	0.80±.02	0.72±.02	0.92±.01	0.90±.02	0.84±.01	0.96±.02	0.86±.03	0.79±.02	0.94±.01	0.79	3.12	0.07	0.90	3.12	0.07	0.86	3.12	0.07
ModaUNETR^E	**0.83**±.03	**0.76**±.02	**0.94**±.01	**0.91**±.02	**0.86**±.01	**0.97**±.01	**0.88**±.03	**0.81**±.02	**0.95**±.02	**0.81**	3.38	0.07	**0.91**	3.38	0.07	**0.87**	3.38	0.07

Table 1: Five-fold cross-validation results on the BraTS 2020 training set for ET, TC, and WT in terms of DSC, IoU, and VS (mean±std). Held-out results on the BraTS 2020 validation set are also reported as DSC, peak GPU memory per volume (GB), and avg inference time (s).

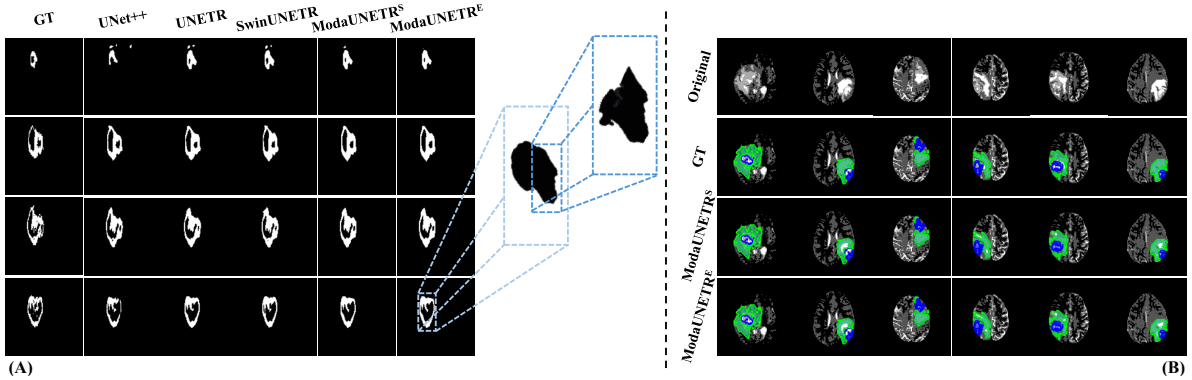

Figure 3: Qualitative comparison of segmentation outputs on representative BraTS 2020 axial slices under five-fold cross-validation. (A) Ground-truth vs. predicted masks. (B) Ground-truth vs. predicted overlaid segmentations. For clarity, only TC (blue) and WT (green) are shown in the 2D overlays. A 3D cutaway rendering is also provided for ModaUNETR^E to highlight tumor structure.

Model	ET	TC	WT
[Wang *et al.*, 2020]	0.79	**0.91**	0.86
[An *et al.*, 2024]	0.72	0.89	0.79
[Kunjumon *et al.*, 2025]	**0.83**	0.89	**0.87**
[Zhou *et al.*, 2025]	0.77	0.90	0.85
ModaUNETR^S (Ours)	0.79	0.90	0.86
ModaUNETR^E (Ours)	0.81	**0.91**	**0.87**

Table 2: DSC comparison with state-of-the-art methods on the BraTS 2020 validation set.

Model	ET	TC	WT	GB	s
UNETR	0.71	0.84	0.78	0.36	0.03
ModaUNETR^S	0.79	0.90	0.86	3.12	0.07
ModaUNETR^E	**0.81**	**0.91**	**0.87**	3.38	0.07

Table 3: Ablation study isolating the effect of structured modality-aware interaction. All models share the same UNETR backbone; results are reported as DSC, memory usage (GB), and inference time (s) on the BraTS 2020 validation set.

Limitations and future work. MATI currently assumes a fixed modality-aligned subspace partition, which fits the BraTS setting where modalities are well-defined and consistently available. In scenarios with missing, corrupted, weakly correlated, or highly redundant modalities, this assumption may be partially violated and may also reduce the interpretability of fixed subspace assignments, motivating adaptive or learned allocation strategies. Moreover, this work is limited to a single backbone (UNETR) and benchmark (BraTS 2020). Since MATI is backbone-agnostic and operates directly on token embeddings, future extensions will validate it across additional architectures and datasets, also investigating adaptive handling of missing or unreliable modalities and dynamic subspace allocation.

6 Conclusions

This work addresses a key limitation of current transformer-based multimodal models: the lack of explicit architectural control over modality interaction in the representation space. We thus introduced MATI, an architecturally lightweight and backbone-agnostic interaction module that injects a structured multimodal inductive bias by factorizing token embeddings into modality-aligned subspaces and regulating cross-modal information exchange. We instantiated MATI within UNETR in two variants: ModaUNETR^S applies structured interaction to selected skip-token representations, whereas ModaUNETR^E injects it directly into the transformer encoder for progressive modality-aware refinement.

Extensive experiments on BraTS 2020 show that both ModaUNETR^S and ModaUNETR^E consistently improve over UNETR under five-fold cross-validation and official, held-out validation, while preserving a single-stream architecture and maintaining a favorable accuracy-efficiency trade-off. More broadly, these results suggest that structured interaction should be treated as a first-class architectural principle in multimodal transformers rather than an emergent property of generic self-attention. Explicitly encoding modality structure within the token embedding space enables more coherent, robust, and efficient multimodal representations.

Acknowledgements

This research was partially funded by the Horizon Europe Research and Innovation Programme under Grant No. 101070537, project *CROSSCON–Cross-Platform Open Security Stack for Connected Devices*. Additional support was provided by the PNRR project *FAIR–Future AI Research* (PE00000013), within the NRRP MUR Programme funded by NextGenerationEU, and by the *AI@TN2.0* project funded by the Autonomous Province of Trento, Italy.

References

[An *et al.*, 2024] Dianlong An, Panpan Liu, Yan Feng, Pengju Ding, Weifeng Zhou, and Bin Yu. Dynamic weighted knowledge distillation for brain tumor segmentation. *Pattern Recognition*, 155:110731, 2024.

[Bakas *et al.*, 2017] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1):1–13, 2017.

[Bakas *et al.*, 2018] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell T Shinohara, Christoph Berger, Sungmin M Ha, Matthew Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432, Cham, 2016. Springer International Publishing.

[Dosovitskiy *et al.*, 2020] 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. *International Conference on Learning Representations (ICLR)*, 2020.

[Fareed *et al.*, 2025] Sidra Fareed, Ding Yi, Babar Hussain, and Subhan Uddin. Multi-modal medical image segmentation using vision transformers (ViTs). *Journal of Biohybrid Systems Engineering*, 1(1):1–21, 2025.

[Hatamizadeh *et al.*, 2022a] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. SwinUNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, Cham, 2022. Springer International Publishing.

[Hatamizadeh *et al.*, 2022b] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584, 2022.

[Isensee *et al.*, 2021] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[Kunjumon *et al.*, 2025] Anila Kunjumon, Chinnu Jacob, and R Resmi. Three-dimensional network with squeeze and excitation for accurate multi-region brain tumor segmentation. *International Journal of Imaging Systems and Technology*, 35(2):e70057, 2025.

[Lin *et al.*, 2023] Jianwei Lin, Jiatai Lin, Cheng Lu, Hao Chen, Huan Lin, Bingchao Zhao, Zhenwei Shi, Bingjiang Qiu, Xipeng Pan, Zeyan Xu, et al. CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Transactions on Medical Imaging*, 42(8):2451–2461, 2023.

[Menze *et al.*, 2015] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Niels Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[Myronenko, 2019] Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 311–320, Cham, 2019. Springer International Publishing.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, Cham, 2015. Springer International Publishing.

[Spyridon *et al.*, 2017] Bakas Spyridon, Akbari Hamed, Sotiras Aristeidis, Bilello Michel, Rozycki Martin, Kirby Justin, Freymann John, Farahani Keyvan, and Davatzikos Christos. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *The Cancer Imaging Archive*, 2017.

[Sun, 2025] Jianfei Sun. MedFusion-TransNet: Multi-modal fusion via transformer for enhanced medical image segmentation. *Frontiers in Medicine*, 12:1557449, 2025.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you

need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wang *et al.*, 2020] Yixin Wang, Yao Zhang, Feng Hou, Yang Liu, Jiang Tian, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-pairing learning for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 230–240. Springer, 2020.

[Xing *et al.*, 2022] Zhaohu Xing, Lequan Yu, Liang Wan, Tong Han, and Lei Zhu. NestedFormer: Nested modality-aware transformer for brain tumor segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 140–150. Springer, 2022.

[Zhang *et al.*, 2022] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022. arXiv preprint.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested U-Net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 3–11. Springer, 2018.

[Zhou *et al.*, 2025] Mingzhe Zhou, Jinbao Li, and Yahong Guo. Multi-level channel-spatial attention and light-weight scale-fusion network (MCSLF-Net): Multi-level channel-spatial attention and light-weight scale-fusion transformer for 3D brain tumor segmentation. *Quantitative Imaging in Medicine and Surgery*, 15(7):6301–6325, 2025.