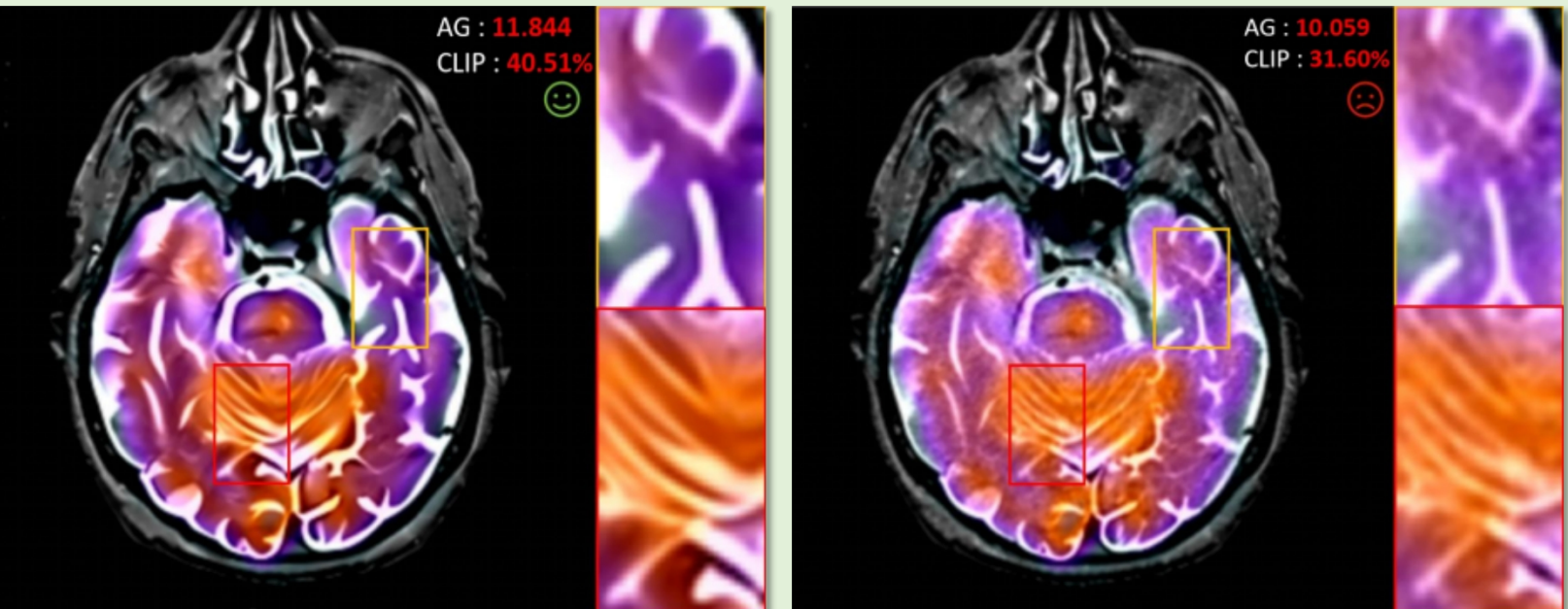


Introduction

Medical Image Fusion (MIF) integrates complementary anatomical and functional information to support clinical diagnosis. However, as shown in existing methods often rely on uniform fusion rules or process-driven instructions, overlooking **pathology-aware diagnostic intent**. Moreover, **Diffusion Transformers (DiTs)** may disrupt anatomical continuity during tokenization and introduce semantic drift during fusion image generation.



Our Goal: To achieve clinically faithful, intent-driven medical image fusion.

Challenges

- Limited diagnostic awareness** under uniform fusion rules.
- Loss of spatial continuity during **DiT tokenization**.
- Semantic-fidelity conflict** during generative optimization.

Quantitative Results

TABLE 1 | FUSION PERFORMANCE ON HARVARD

Higher is better ↑ | Green-filled cells denote the best values in the full original comparison.

CT-MRI					PET-MRI					SPECT-MRI				
Method	EN ↑	MI ↑	SF ↑	SD ↑	Method	EN ↑	MI ↑	SF ↑	SD ↑	Method	EN ↑	MI ↑	SF ↑	SD ↑
DDFM	4.195	2.675	22.165	62.305	DDFM	4.922	2.455	23.376	78.061	DDFM	4.392	3.262	17.399	64.194
FILM	3.907	2.986	42.957	83.367	FILM	3.437	3.746	31.297	72.937	FILM	3.776	2.807	22.687	66.038
Text-DiFuse	5.577	3.247	34.606	87.012	Text-DiFuse	6.224	3.127	35.672	97.248	Text-DiFuse	5.895	3.074	26.081	77.622
TextFusion	4.775	2.944	22.424	81.132	TextFusion	5.770	3.345	32.018	83.078	TextFusion	5.032	3.444	20.240	60.185
MIND (Ours)	5.802	3.274	41.250	92.291	MIND (Ours)	6.238	4.063	42.302	97.716	MIND (Ours)	6.236	4.142	23.316	79.905

Selected SOTA baselines • Mean values

KEY RESULT MIND ranks first in EN, MI and SD across all three modality pairs.

TABLE 2 | INFERENCE EFFICIENCY

Method	Params (M) ↓	Time (s) ↓	FLOPs (G) ↓
DDFM	552.814	11.818	2293.76
Text-DiFuse	119.460	2.135	18180.07
DiTFuse★	37.750	13.370	6469.60
MIND★ (Ours)	39.930	7.851	6860.69

★ LoRA fine-tuning | Best values in green; MIND runner-up in pale yellow.

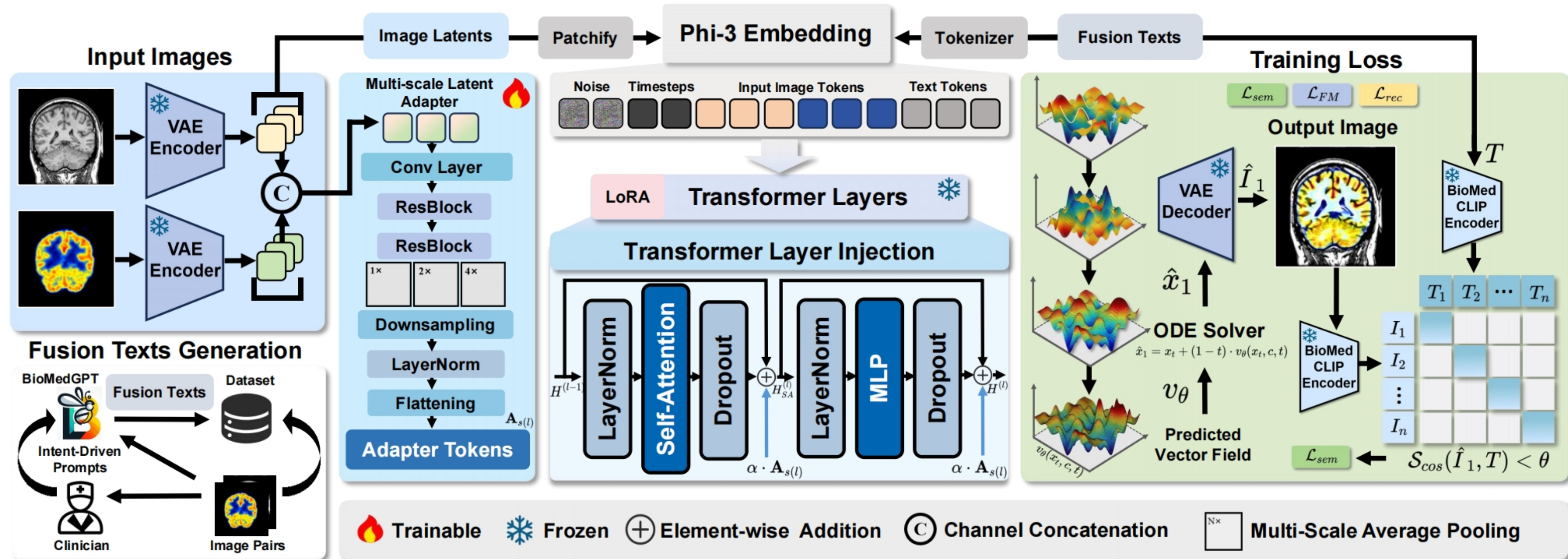
TABLE 3 | BraTS SEGMENTATION

Method	ED ↑	NET ↑	ET ↑	Mean Rank ↓
MR-FLAIR	0.785	0.482	0.450	7.333
MR-T1CE	0.701	0.728	0.588	4.333
DDFM	0.760	0.712	0.552	4.000
MACTFusion	0.757	0.705	0.568	4.333
TextFusion	0.715	0.731	0.579	4.000
SAFusion	0.687	0.646	0.559	7.000
MIND (Ours)	0.786	0.706	0.616	2.000

Original Mean Rank across all methods (including omitted baselines).

MIND 39.93M parameters | 7.851 s/image | BraTS best ED / ET Dice: 0.786 / 0.616 | Mean Rank: 2.000

Methodology



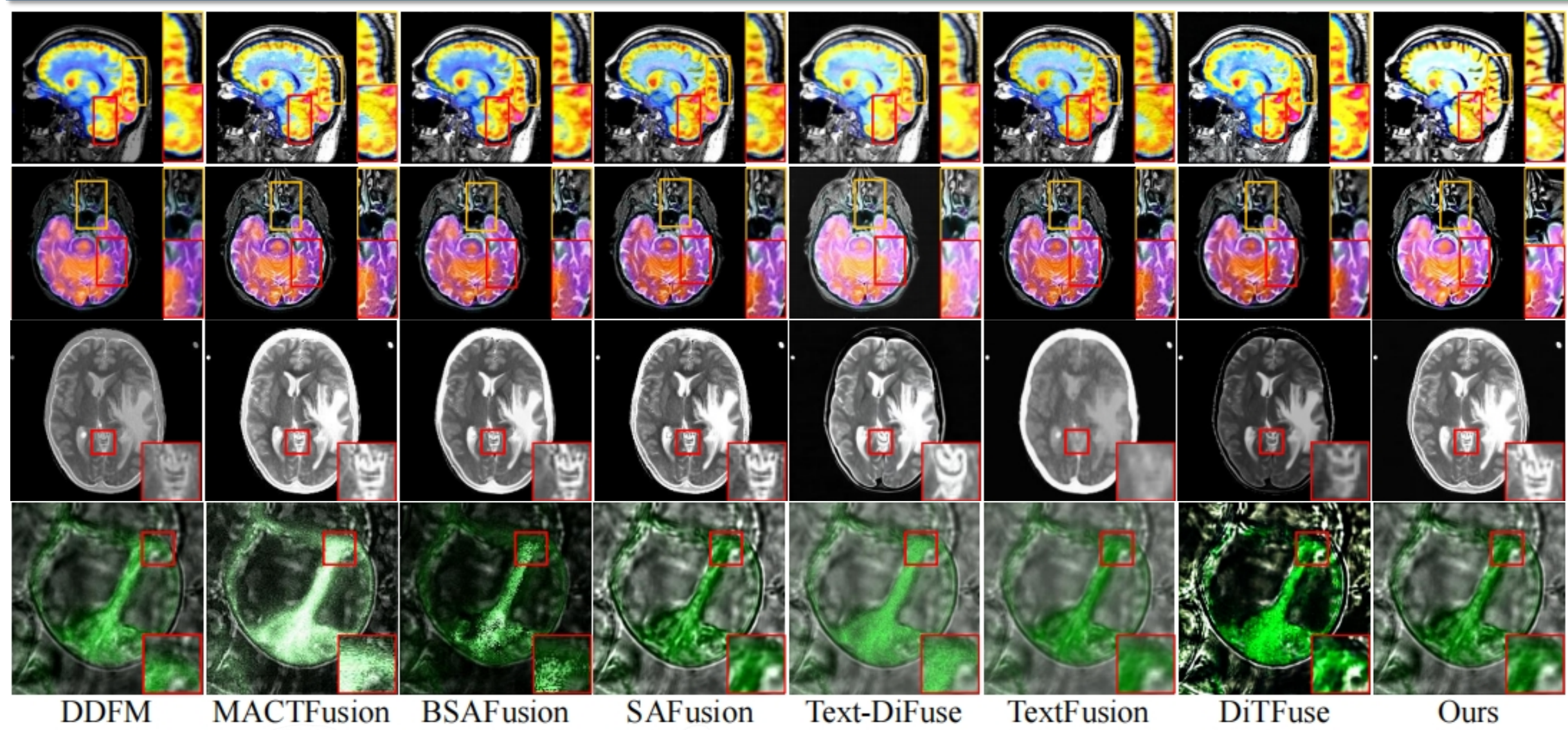
1. We introduce **Intent-Driven Fusion Texts**, leveraging **BioMedGPT** to provide pathology-aware diagnostic guidance.

2. Our **Multi-scale Latent Adapter (MLA)** injects anatomical features into DiTs, preserving spatial continuity and structural details.

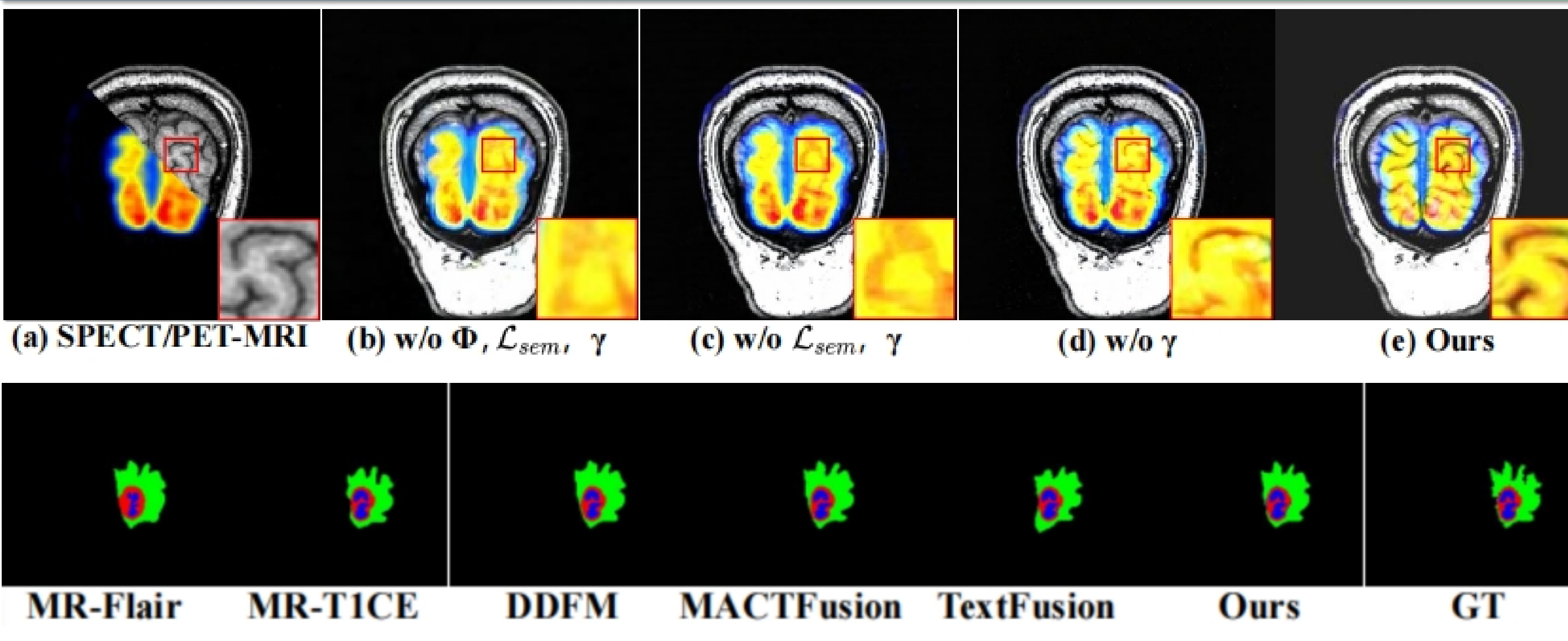
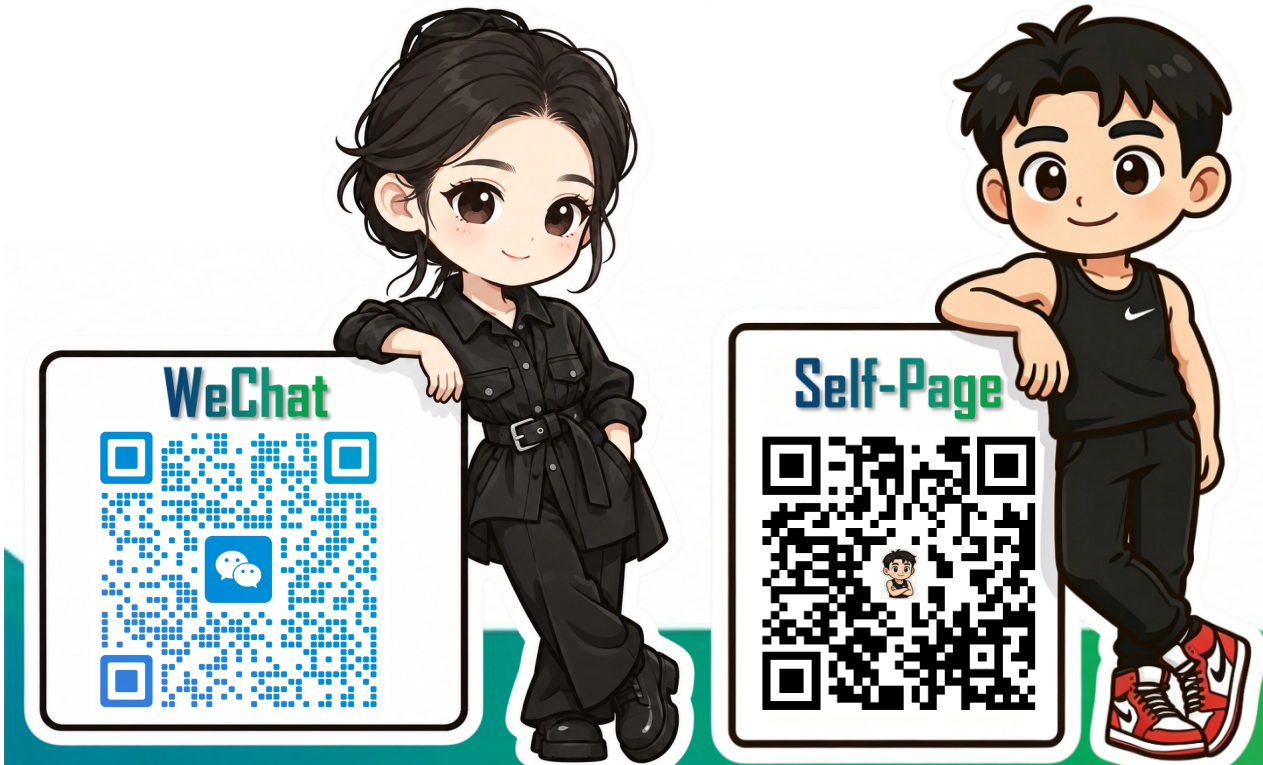
3. We propose a **Timestep-Truncated Semantic Consistency Loss** to align clinical intent while maintaining physical fidelity.

4. Extensive experiments on **Harvard, BraTS, and GFP** demonstrate superior fusion quality, downstream utility, and generalization.

Qualitative Results



Ablation and Downstream Validation



Yunzhan Fu | Email: yunzhanfu24@gmail.com
Jian Wu | Email: wujian2000@zju.edu.cn
Hongxia Xu | Email: einstein@zju.edu.cn