Speech-Guided Multimodal Learning for Vocal Tract Segmentation in Real-Time MRI

Daiqi Liu[1](✉), Lukas Mulzer[1], Md Hasan[1], Nyvenn Alves de Castro[2], Fangxu Xing[3], Xingjian Kang[4], Chengze Ye[1], Siyuan Mei[1], Yipeng Sun[1], Tomás Arias-Vergara[1,6], Jana Hutter[5], Jonghye Woo[3], Andreas Maier[1], and Paula Andrea Pérez-Toro[1,6]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
daiqi.deutschfau.liu@fau.de
[2] Smart Imaging Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[3] Harvard Medical School / Massachusetts General Hospital, Boston, USA
[4] Center for AI and Data Science, Julius-Maximilians-Universität Würzburg, Würzburg, Germany
[5] Institute for Information Processing, Leibniz University Hannover, Hannover, Germany
[6] GITA Lab, Facultad de Ingeniería. Universidad de Antioquia UdeA, Medellín, Colombia

Abstract. Segmenting vocal tract articulators in real-time MRI (rtMRI) is a challenging dynamic image segmentation problem characterized by low contrast, rapid motion, and limited spatial resolution. However, while rtMRI acquisitions may provide synchronized acoustic signals, existing methods discard this information, and the few multimodal approaches that incorporate audio cannot be deployed when audio is unavailable. We propose a three-stage framework that leverages acoustic and phonological supervision during training while requiring only the rtMRI image at inference: phonological representations are converted into spatial bounding-box priors for articulator localization, visual and acoustic encoders are aligned via dual-level cross-modal contrastive pretraining, and the learned representations are fused through a cross-attention decoder, effectively transferring multimodal knowledge into a single-modality inference pipeline. Evaluated on 75-Speaker Annot-16 and USC-TIMIT datasets, our method outperforms existing unimodal and multimodal methods, demonstrating that multimodal supervision provides transferable benefits for precise and clinically deployable vocal tract segmentation. Our code is available at https://github.com/daiqi76/SGVocSegMRI.

Keywords: Segmentation · Multimodal Learning · Real-time MRI · Vocal Tract.

1 Introduction

Real-time MRI (rtMRI) enables simultaneous, non-invasive visualization of all vocal tract articulators during continuous natural speech, making it an indispensable tool for both phonetic research and clinical speech analysis [31]. Accurate segmentation of these structures is critical for a wide range of clinical and translational applications, including pre-surgical planning for glossectomy, assessment of structural speech disorders (e.g., cleft lip and palate), monitoring articulatory decline in neurodegenerative diseases such as ALS, and constructing articulatory representations for speech neuroprosthetics and rehabilitation [2, 8, 15].

Significant progress in medical image segmentation has been driven by deep learning, from the seminal U-Net [27] and its variants to Transformer-based architectures such as Swin-UNETR [9] and vision foundation models including SAM [13] and its medical adaptations [18]. State space models such as U-Mamba have further extended sequential modeling capabilities to volumetric and high-resolution medical images [19]. Within the specific domain of vocal tract segmentation, early efforts to extract articulatory contours largely relied on manual or semi-automated boundary tracing, where air–tissue boundaries are hand-annotated in a reference frame, followed by nonlinear optimization propagating contours across subsequent frames [26]. Other approaches assigned labels by examining pixel intensities along predefined gridlines overlaid on MR images, or by employing active shape models to capture the expected anatomical contours [14]. This process is not only time-consuming and labor-intensive but also susceptible to error, requiring extensive human supervision. Subsequent deep learning approaches applied fully convolutional networks to air-tissue boundary labeling [20, 30] and U-Net-based architectures to multi-structure segmentation, achieving a Dice score of 0.85 [28]. Most recently, a multimodal framework combining rtMRI with synchronized speech audio via Transformer-based fusion established a new benchmark in speaker-independent vocal tract segmentation [12].

Despite these advances, existing multimodal approaches rely on implicit feature concatenation without explicitly modeling the correspondence between visual and acoustic representations, overlooking the complementary role of phonological class features. More critically, these methods require audio at inference time, yet clinical deployment frequently faces modality absence: conventional metallic microphones cannot be used inside MRI scanners, while fiber-optic alternatives remain prohibitively expensive; furthermore, target patient populations such as those with tongue cancer or dysarthria often present with severe speech disorders, making clean audio acquisition unreliable or impossible.

To address these limitations, we propose a three-stage multimodal framework that integrates visual, acoustic, and phonological information for precise articulator segmentation in rtMRI. Stage 1 converts phonological class descriptors into subject-specific spatial bounding-box priors, providing explicit localization guidance for small and low-contrast structures. Stage 2 pretrains the visual and audio encoders via a dual-level cross-modal contrastive objective, establishing a shared semantic space through both global and local alignment. Stage 3 fuses the aligned representations through a cross-attention decoder, in which visual

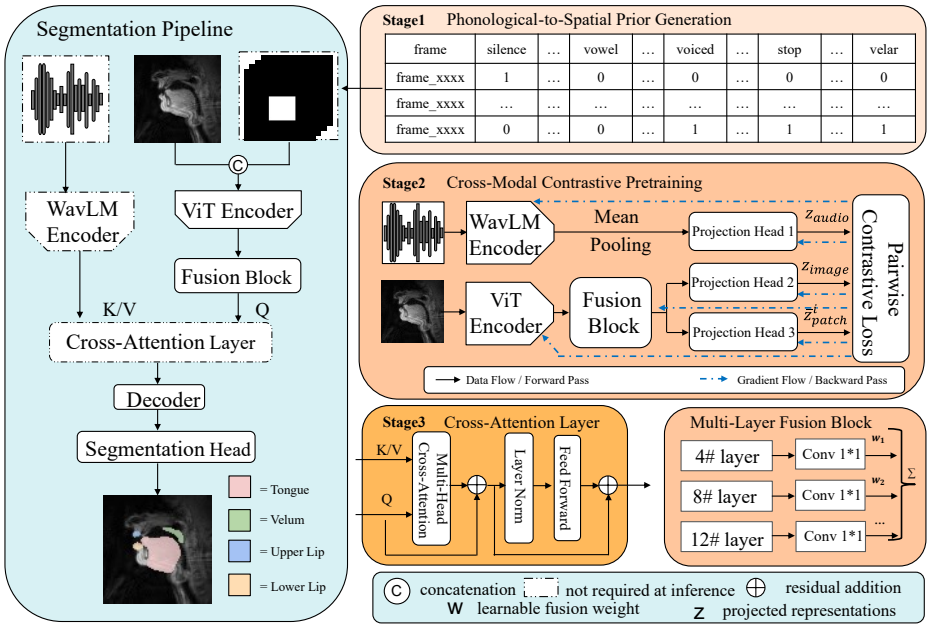

Fig. 1. Schematic overview of the proposed multimodal segmentation framework. Left: The segmentation pipeline operates with three input modalities during training (rtMRI image, audio, and phonological bounding-box prior) but requires only the image at inference time. Right: Three training stages are illustrated from top to bottom.

tokens dynamically attend to temporally resolved audio features for fine-grained multimodal integration. Critically, the learned cross-modal representations are fully encoded into the model weights at training time, such that only the rtMRI image is required at inference, which makes our method directly deployable in clinical settings where audio may be absent or degraded. Experiments on the 75-Speaker Annot-16 [29] and USC-TIMIT [22] datasets demonstrate state-of-the-art performance, confirming that structured phonological supervision and explicit cross-modal alignment provide transferable benefits that persist at inference even without audio input.

2 Methods

As illustrated in Fig. 1, Our method is structured into three stages that progressively incorporate phonological, acoustic, and visual information to achieve precise articulator delineation in rtMRI. Crucially, only the image is required at inference.

2.1 Stage 1: Phonological-to-Spatial Prior Generation

Phonological descriptors encode structured linguistic knowledge about articulatory configurations across three dimensions: **voicing** (whether vocal folds vibrate during sound production), **manner of articulation** (how airflow is shaped in the vocal tract), and **place of articulation** (where in the vocal tract the constriction occurs), providing symbolic categorical representations that explicitly describe how speech sounds are produced. For instance, the phoneme /p/ can be characterized as [voiceless, stop, labial], directly indicating the involvement of lip closure. Rather than incorporating these descriptors directly as text-based semantic features, given a multi-hot encoded phonological descriptor $\mathbf{P} \in \{0,1\}^{15}$, we generate a four-channel spatial prior by mapping each phonological class to the bounding box of its associated articulatory region. Formally, for each articulator channel c, the prior is computed as:

$$\mathbf{x}_{\text{bbox}}^{(c)} = \text{BBox}\Big(\bigcup_{i:\, p_i^{(c)}=1} \mathbf{S}_i^{(c)}\Big) \in \mathbb{R}^{H\times W} \tag{1}$$

where $\mathbf{S}_i^{(c)}$ denotes the ground-truth segmentation mask of articulator c in frame i, and the union is taken over all frames sharing the same phonological attribute. The operator $\text{BBox}(\cdot)$ encloses the union region with its minimum bounding rectangle, yielding an interpretable spatial prior for each of the four articulators (tongue, velum, upper lip, lower lip). To accommodate inter-speaker anatomical variability, priors are computed on a subject-specific basis.

2.2 Stage 2: Cross-Modal Contrastive Pretraining

Given a synchronized pair $\{\mathbf{I}, \mathbf{A}\}$, where $\mathbf{I} \in \mathbb{R}^{1\times H\times W}$ is an rtMRI frame and $\mathbf{A} \in \mathbb{R}^{T_a}$ is the aligned audio segment, we extract modality-specific features using two pretrained encoders: ViT-Base [4, 32] (E_I) for visual input and WavLM-Base [3] (E_A) for audio, both with hidden dimension D=768. From E_I, patch tokens (excluding the `[CLS]` token) from layers $\{4, 8, 12\}$ are fused via learnable 1×1 convolutions to obtain multi-scale local features.

$$\mathbf{Z}_{\text{patch}} = \mathcal{R}^{-1}\Big(\textstyle\sum_{l\in\{4,8,12\}} \mathbf{W}_l * \mathcal{R}\Big(E_I^{(l)}(\mathbf{I})[1{:}]\Big)\Big) \in \mathbb{R}^{N\times D} \tag{2}$$

where $\mathcal{R}(\cdot) : \mathbb{R}^{N\times D} \to \mathbb{R}^{D\times \frac{H}{P}\times \frac{W}{P}}$ denotes the reshape operation converting patch tokens to spatial feature maps, $\mathcal{R}^{-1}(\cdot)$ is its inverse. The final-layer `[CLS]` token serves as the global visual representation $\mathbf{Z}_{\text{image}} \in \mathbb{R}^D$. Concurrently, the audio feature sequence is mean-pooled along the temporal dimension to obtain a compact global representation. All representations are projected into a shared 256-dimensional space via two-layer MLPs. We enforce dual-level alignment using the InfoNCE loss [23]. Global-to-global alignment matches `[CLS]` tokens with audio ($\mathcal{L}_{\text{g2g}} = \mathcal{L}(\mathbf{z}_{\text{image}}, \mathbf{z}_{\text{audio}})$), while local-to-global alignment enforces consistency between mean-pooled patch features and audio ($\mathcal{L}_{\text{l2g}} = \mathcal{L}(\bar{\mathbf{z}}_{\text{patch}}, \mathbf{z}_{\text{audio}})$). The total contrastive loss is defined as $\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{g2g}} + \lambda\mathcal{L}_{\text{l2g}}$, where $\lambda = 0.5$.

2.3 Stage 3: Cross-Attention-Based Multimodal Segmentation

The rtMRI frame $\mathbf{I}$ and the phonological bounding-box map $\mathbf{x}_{\text{bbox}}$ from Stage 1 are concatenated channel-wise and processed through the Stage 2 pretrained ViT encoder and fusion block to obtain visual patch tokens $\mathbf{F}_v \in \mathbb{R}^{N\times D}$. Concurrently, WavLM encodes the audio into a full temporal sequence $\mathbf{F}_a \in \mathbb{R}^{T\times D}$ without pooling. A six-layer cross-attention decoder refines visual tokens by attending to audio features at each layer:

$$\mathbf{F}_v^{(\ell)} = \text{LayerNorm}\left(\mathbf{F}_v^{(\ell-1)} + \text{MultiHeadAttn}\left(\mathbf{F}_v^{(\ell-1)}, \mathbf{F}_a, \mathbf{F}_a\right)\right), \tag{3}$$

where visual tokens serve as queries and audio tokens as keys and values. The refined tokens are upsampled through two convolutional blocks (768→256→128 channels) with bilinear interpolation (14×14→224×224), followed by a 1×1 convolution yielding class logits $\mathbf{Y} \in \mathbb{R}^{4\times H\times W}$. Training minimizes the combined loss $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}}(\mathbf{Y}, \mathbf{S}_{\text{gt}}) + \mathcal{L}_{\text{Dice}}(\mathbf{Y}, \mathbf{S}_{\text{gt}})$.

3 Experiments

3.1 Dataset & Preprocessing

75-Speaker Annot-16 This dataset provides synchronized rtMRI videos and audio recordings from 16 speakers (7F, 9M; 8 native and 8 non-native American English speakers), with phonetic alignments and handmade articulator contour annotations. The rtMRI was acquired at 83.28 fps with 2.4 mm in-plane resolution (84×84 pixels) using a 1.5 T scanner; audio was recorded at 20 kHz via a fiber-optic microphone. We use 12 speakers for training (51,970 frames) and 5 held-out speakers for evaluation, organized into three configurations: Seen Speaker, Unseen Task (SS-UT), Unseen Speaker, Seen Task (US-ST), and Unseen Speaker, Unseen Task (US-UT) (9,148 frames).

USC-TIMIT This dataset comprises recordings from 10 native American English speakers (5M, 5F), each producing 460 phonetically balanced sentences, acquired at 23.18 fps with 2.9 mm isotropic resolution (68×68 pixels) using a 1.5 T scanner. Synchronized speech was simultaneously recorded using a fiber-optic microphone. Articulator contours were extracted using the semi-automatic segmentation tool provided with the dataset. We select two held-out speakers (1M, 1F) for evaluation under the US-UT configuration (5,731 frames).

Preprocessing We apply the same pipeline to both datasets. The rtMRI videos were temporally downsampled to 15 fps, from which individual frames were extracted, resized to 224×224 pixels via bilinear interpolation, and normalized per subject to $[0, 1]$ based on subject-specific minimum and maximum intensity values. Synchronized audio was resampled to 16 kHz. For each extracted frame, a three-window audio context was constructed by concatenating the aligned central segment with its preceding and following windows. Phoneme labels are mapped to 15-dimensional multi-hot phonological feature vectors [1, 16], and articulator contours were converted to pixel-level binary masks and reviewed by a speech expert, yielding four articulator classes: tongue, velum, upper lip, and lower lip.

Table 1. Performance comparison on 75-Speaker Annot-16 and USC-TIMIT datasets. The best results are highlighted in **bold**, and the second-best are underlined.

Methods	75-Speaker Annot-16						USC-TIMIT	
	SS-UT		US-ST		US-UT		US-UT	
	DSC(%)↑	ASD(mm)↓	DSC(%)↑	ASD(mm)↓	DSC(%)↑	ASD(mm)↓	DSC(%)↑	ASD(mm)↓
State-of-The-Art Unimodal Methods								
nnU-Net [11]	74.59 ± 10.09	5.66 ± 4.79	73.11 ± 11.56	5.90 ± 7.67	69.77 ± 10.49	6.87 ± 8.58	66.55 ± 13.27	7.81 ± 9.35
ResUNet-a [5]	69.24 ± 12.14	6.13 ± 5.27	68.27 ± 10.43	6.60 ± 5.77	67.58 ± 11.27	7.41 ± 9.35	65.67 ± 12.78	7.66 ± 8.29
Swin-UNETR [9]	72.38 ± 10.58	5.22 ± 4.91	72.72 ± 9.11	5.06 ± 6.98	70.61 ± 10.93	4.37 ± 6.20	69.40 ± 11.19	4.86 ± 7.54
MedSAM-2 [34]	75.14 ± 8.93	3.11 ± 3.96	73.29 ± 8.57	3.17 ± 4.10	71.57 ± 9.12	3.50 ± 4.51	69.22 ± 9.61	4.48 ± 5.46
MedSegDiff-V2 [33]	69.01 ± 12.47	5.10 ± 3.44	67.04 ± 18.18	5.08 ± 3.14	65.43 ± 13.09	6.71 ± 5.17	63.64 ± 13.28	5.59 ± 4.20
DINOv2 [24]	77.83 ± 11.53	3.76 ± 4.55	77.70 ± 9.34	3.51 ± 4.10	73.58 ± 10.84	3.50 ± 5.11	71.38 ± 11.66	4.42 ± 5.71
U-Mamba [19]	81.71 ± 8.70	2.52 ± 3.48	80.90 ± 8.26	2.98 ± 3.77	79.49 ± 9.17	3.82 ± 4.83	78.06 ± 10.47	4.11 ± 4.53
State-of-The-Art Multimodal Methods								
AVSegFormer* [7]	83.68 ± 7.94	2.80 ± 4.68	81.27 ± 6.40	$\underline{2.39 \pm 4.32}$	80.53 ± 8.73	$\underline{3.48 \pm 4.77}$	76.59 ± 9.16	$\mathbf{3.02 \pm 4.93}$
AV-SAM* [21]	82.27 ± 7.14	2.38 ± 2.83	83.47 ± 7.64	3.01 ± 2.19	$\underline{81.21 \pm 8.52}$	3.69 ± 3.49	$\underline{80.62 \pm 8.23}$	4.26 ± 4.88
Jain et al. [12]	$\underline{85.35 \pm 7.25}$	$\underline{2.21 \pm 3.76}$	$\underline{84.68 \pm 8.33}$	2.50 ± 3.58	81.19 ± 9.36	5.54 ± 3.12	80.14 ± 10.42	3.50 ± 3.41
Ours	$\mathbf{90.82 \pm 6.30}$	$\mathbf{1.48 \pm 1.72}$	$\mathbf{87.74 \pm 6.41}$	$\mathbf{2.11 \pm 2.43}$	$\mathbf{86.37 \pm 7.04}$	$\mathbf{2.36 \pm 2.90}$	$\mathbf{83.78 \pm 8.25}$	$\underline{3.43 \pm 3.17}$

*Adapted.

3.2 Implementation Details

All experiments are implemented in PyTorch 2.5.1 [25] and conducted on an NVIDIA RTX A100 GPU with a batch size of 8. In Stage 2, the model is trained for 50 epochs using AdamW with weight decay 1×10^{-2}. During phase 1 (epochs 1-20), both encoders are frozen and only the fusion block and projection heads are optimized at a learning rate of 1×10^{-4}. During phase 2 (epochs 21–50), ViT layers are progressively unfrozen from top to bottom: Layer 12 at epoch 21, Layers 9–12 at epoch 31, each with a reduced learning rate of 1×10^{-5} to preserve pretrained representations. In Stage 3, the pretrained ViT encoder, WavLM encoder, and fusion block are frozen to preserve the cross-modal representations learned in Stage 2. Only the newly introduced cross-attention layers and segmentation decoder are optimized using AdamW with a learning rate of 1×10^{-4} and weight decay 1×10^{-2}. The segmentation loss combines cross-entropy and Dice losses with equal weights. Training proceeds for up to 100 epochs with early stopping (patience = 15) based on validation Dice score.

3.3 Performance Comparison

We compare our method with state-of-the-art segmentation models, including both unimodal methods that rely solely on visual input and multimodal approaches. Evaluation metrics include Dice Coefficient (DSC) and Average Surface Distance (ASD). As shown in Tab. 1, our method achieves leading performance on both datasets. On the most challenging (US-UT) setting, ours achieves further gains of +6.88 DSC, and −1.46 mm ASD; similar improvements are observed on USC-TIMIT (+5.72 DSC). Notably, ours also outperforms recent foundation models including DINOv2 [24] and MedSAM-2 [34] by large margins. Among multimodal methods, ours consistently outperforms all competitors, achieving +5.18 DSC and −3.18 mm ASD over the strongest competitor Jain et al. [12] on the US-UT configuration. On USC-TIMIT, although AVSegFormer achieves

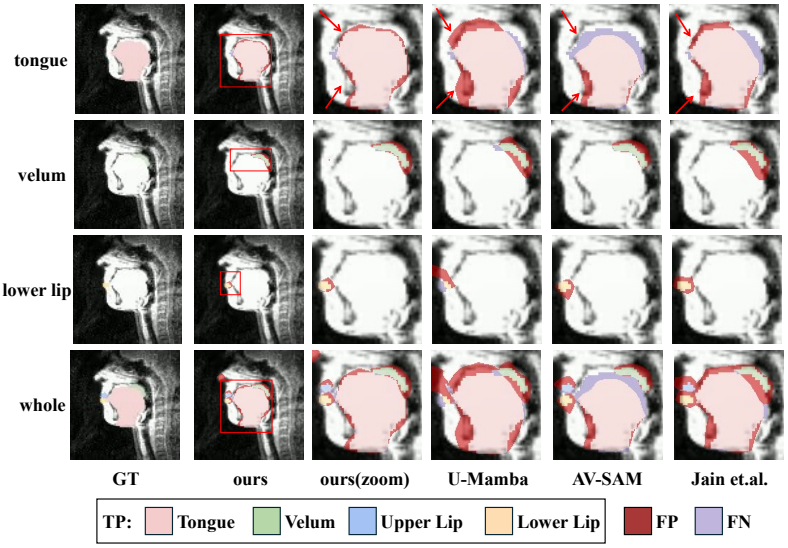

Fig. 2. Qualitative comparison of articulator segmentation on a representative rtMRI frame. Colored regions in the figure indicate true positives for each articulator class. Red regions denote False Positive (FP), and purple regions denote False Negatives (FN). Red arrows highlight failure cases in competing methods. TP: True Positive.

a marginally lower ASD, our method exhibits substantially smaller standard deviation, indicating more consistent and stable segmentation across subjects.

Visually, as Fig. 2 shows, for the tongue (row 1), competing methods exhibit clear over-segmentation artifacts at both the superior and inferior boundaries (indicated by red arrows), incorrectly extending the predicted mask beyond the true tongue contour. In contrast, Our method closely adheres to the ground truth boundary. Ours also shows clear advantages in segmenting challenging small structures, such as the lower lip (row 3), yielding the lowest number of both false-positive and false-negative predictions.

3.4 Ablation Study

As shown in Tab. 2 (A), we conduct ablation studies to verify the effectiveness of each proposed stage. We adopt the architecture of Jain et al. [12] as baseline, which employs feature concatenation for cross-modal fusion. Introducing the audio modality IA Concat Fusion already yields a notable improvement of +3.92 DSC and −0.87 mm ASD over Image-only, confirming the complementary role of acoustic information in articulator segmentation. However, naively appending phonological features via concatenation IAP Concat Fusion provides only marginal gains (+0.35 DSC). In contrast, converting phonological descriptors into spatial bounding-box priors +BBox Prior yields a substantial improvement of +2.23 DSC and −2.15 mm ASD, demonstrating the effectiveness

Table 2. Ablation study of our method on the 75-Speaker Annot-16 dataset under the US-UT configuration. (A) Effect of each stage: "Image-only" denotes the baseline using image input solely. Inference time is measured in ms per 50 frames.(B) Effect of different image encoders. The best results are highlighted in **bold**, and the second-best are underlined. I: Image. A: Audio. Phon and P: Phonological class.

(A) Ablation Study of Different Stages

Method				Input (Inference)			DSC(%)↑	ASD(mm)↓	#Params(M)	Time(ms)↓
Name	Stage 1	Stage 2	Stage 3	Image	Audio	Phon				
Image-only	×	×	×	✓	×	×	77.29 ± 14.59	6.41 ± 5.22	**109.77**	**510**
IA Concat Fusion	×	×	×	✓	✓	×	81.21 ± 9.36	5.54 ± 3.12	161.53	1173
IAP Concat Fusion	×	×	×	✓	✓	✓	81.56 ± 8.71	6.03 ± 3.61	162.27	1214
+ BBox Prior	✓	×	×	✓	×	×	83.79 ± 10.31	3.88 ± 4.51	161.53	1008
+ Pretrain	✓	✓	×	✓	×	×	84.70 ± 9.18	3.64 ± 3.77	163.10	965
w/ full input	✓	✓	×	✓	✓	✓	$\mathbf{86.59 \pm 6.32}$	$\mathbf{2.14 \pm 3.13}$	186.64	1356
ours	✓	✓	✓	✓	×	×	86.37 ± 7.04	2.36 ± 2.90	186.64	1058

(B) Ablation Study of Image Encoder

Image Encoder	DSC(%)↑	ASD(mm)↓	#Params(M)	Time(ms)↓
ResNet-34 [10]	82.26 ± 5.97	5.11 ± 3.28	**121.25**	**751**
Swin-B [17]	83.11 ± 9.51	4.36 ± 4.14	188.92	1245
DINOv2-B [24]	84.03 ± 8.43	3.55 ± 4.03	186.64	1072
ViT-B (Ours) [6]	$\mathbf{86.37 \pm 7.04}$	$\mathbf{2.36 \pm 2.90}$	186.64	1058

of Stage 1 in bridging the modality gap. Progressively incorporating Stage 2 contrastive pretraining +Pretrain and Stage 3 cross-attention fusion brings further consistent gains across all metrics. Critically, comparing w/ full input (1356 ms) and ours (1058 ms) reveals that image-only inference reduces latency by 22% while incurring only a marginal performance gap (−0.22 DSC, +0.22 mm ASD), demonstrating that the cross-modal knowledge acquired during training is effectively distilled into the model weights. Tab. 2 (B) presents the results of replacing the image encoder with alternative architectures while keeping all other components fixed. Although ResNet-34 [10] achieves the fastest inference time (751 ms) and reduces the total parameter count by 65.39M, it yields a notable performance drop of −4.11 DSC and +2.75 mm ASD compared to ViT-B [6]. Swin-B [17], DINOv2-B [24] achieve a similar parameter count, but the model with ViT-B encoder achieves the best trade-off between performance, parameter efficiency, and inference speed.

4 Discussion & Conclusion

We present the first framework to combine visual and acoustic information for vocal tract articulator segmentation in rtMRI while requiring only image modality at inference time. By explicitly incorporating phonological class priors as spatial bounding-box maps, our method bridges the modality gap between symbolic linguistic knowledge and spatial image representations. By pretraining the encoders via contrastive learning, the learned cross-modal representations are fully encoded into the model weights, making the framework robust to audio absence in clinical deployment. Furthermore, audio-guided cross-attention allows visual tokens to dynamically attend to temporally resolved acoustic features, enabling fine-grained integration of articulatory motion patterns. These results

demonstrate that structured phonological supervision and cross-modal alignment provide complementary and transferable benefits that persist at inference even without audio input.

Nevertheless, there is still room for improvement. The current framework processes each frame independently, without modeling temporal dependencies across frames, which may introduce jitter in predicted trajectories. Furthermore, phonological features are treated as context-free labels, neglecting coarticulation effects that systematically shape articulator positions based on neighboring phonemes. Finally, all experiments are conducted on healthy speaker data due to the scarcity of annotated pathological rtMRI datasets. Future work should explore temporal architectures, context-aware phoneme embeddings derived from speech language models, and validation on dysarthric and post-surgical speech data to fully assess the clinical utility of the proposed framework.

Acknowledgments. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

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

References

1. Arias-Vergara, T., et al.: Contrastive learning approach for assessment of phonological precision in patients with tongue cancer using mri data. In: Interspeech. p. 927 (2024)
2. Browman, C.P., et al.: Articulatory phonology: An overview. Phonetica **49**(3-4), 155–180 (1992)
3. Chen, S., et al.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (2022)
4. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)
6. Dosovitskiy, A., Beyer, L., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 12155–12163 (2024)
8. Hagedorn, C., et al.: Characterizing post-glossectomy speech using real-time mri. In: International Seminar on Speech Production, Cologne, Germany. pp. 170–173 (2014)
9. Hatamizadeh, A., et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)

10. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Isensee, F., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Jain, R., et al.: Multimodal segmentation for vocal tract modeling. Interspeech (2024)
13. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Labrunie, M., et al.: Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning. Speech Communication **99**, 27–46 (2018)
15. Lammert, A.C., et al.: Investigation of speed-accuracy tradeoffs in speech production using real-time magnetic resonance imaging. In: Proc. Interspeech 2016. pp. 460–464 (2016)
16. Liu, D., et al.: Audio–vision contrastive learning for phonological class recognition. In: International Conference on Text, Speech, and Dialogue. pp. 60–71. Springer (2025)
17. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. Ma, J., et al.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
19. Ma, J., et al.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
20. Mannem, R., et al.: Air-tissue boundary segmentation in real time magnetic resonance imaging video using a convolutional encoder-decoder network. In: ICASSP. pp. 5941–5945. IEEE (2019)
21. Mo, S., Tian, Y.: Av-sam: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023)
22. Narayanan, S., et al.: Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). The Journal of the Acoustical Society of America **136**(3), 1307–1311 (2014)
23. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
24. Oquab, M., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
25. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
26. Ramanarayanan, V., et al.: Analysis of speech production real-time mri. Computer Speech & Language **52**, 1–22 (2018)
27. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
28. Ruthven, M., et al.: Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech. Computer Methods and Programs in Biomedicine **198**, 105814 (2021)
29. Shi, X., et al.: 75-speaker annot-16: A benchmark dataset for speech articulatory rt-mri annotation with articulator contours and phonetic alignment. Proc. Interspeech 2025 pp. 2175–2179 (2025)

30. Somandepalli, K., et al.: Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images. In: Interspeech. pp. 631–635 (2017)
31. Toutios, A., et al.: Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research. APSIPA Transactions on Signal and Information Processing **5**, e6 (2016)
32. Wu, B., et al.: Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
33. Wu, J., et al.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6030–6038 (2024)
34. Zhu, J., et al.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)