Harnessing Text Insights with Visual Alignment for Medical Image Segmentation

Qingjie Zeng, Huan Luo, Zilin Lu, Yutong Xie, Zhiyong Wang, *Member, IEEE*, Yanning Zhang, *Fellow, IEEE*, and Yong Xia, *Member, IEEE*

***Abstract*— Pre-trained vision-language models (VLMs) and language models (LMs) have recently garnered significant attention due to their remarkable ability to represent textual concepts, opening up new avenues in vision tasks. In medical image segmentation, efforts are being made to integrate text and image data using VLMs and LMs. However, current text-enhanced approaches face several challenges. First, using separate pre-trained vision and text models to encode image and text data can result in semantic shifts. Second, while VLMs can establish the correspondence between visual and textual features when pre-trained on paired image-text data, this alignment often deteriorates during segmentation tasks due to misalignment between the text and vision components in ongoing learning. In this paper, we propose TeViA, a novel approach that seamlessly integrates with various vision and text models, irrespective of their pre-training relationships. This integration is achieved through a segmentation-specific text-to-vision alignment design, ensuring both information gain and semantic consistency. Specifically, for each training data, a foreground visual representation is extracted from the segmentation head and used to supervise projection layers, thereby adjusting the textual features to better contribute to the segmentation task. Additionally, a historic visual prototype is created by aggregating target semantics from all training data and is updated using a momentum-based manner. This prototype aims to enhance the visual representation of each data instance by establishing feature-level connections, which in turn refines the textual features. The superiority of TeViA is validated on five public datasets, exhibiting over 6% Dice improvements compared to vision-only methods. Code is available at: https://github.com/jgfiuuuu/TeViA.**

***Index Terms*— Medical image segmentation, textual knowledge, visual alignment**

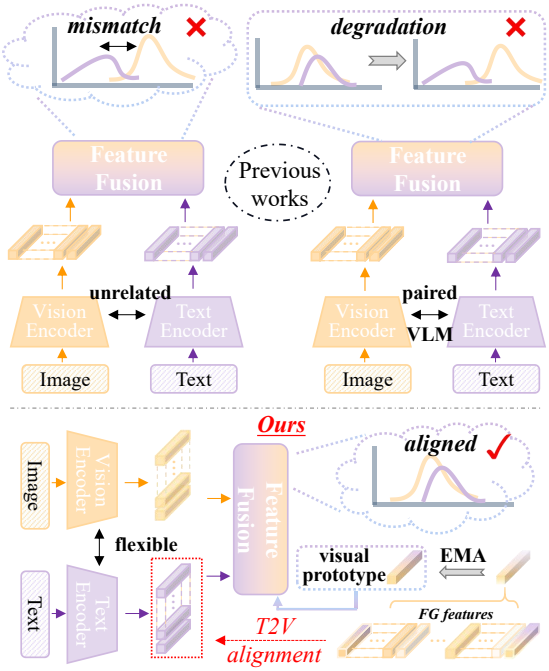


Fig. 1. **Left Top:** This part presents methods that use independently pre-trained vision and text encoders for segmentation, which often face the challenge of mismatched feature distribution. **Right Top:** This section shows methods that leverage VLMs pre-trained on paired image-text data for segmentation. Although textual and visual features are initially aligned, this correspondence can easily be disrupted when performing segmentation with only visual supervision. **Bottom:** Our method seamlessly integrates with various vision and text encoders, whether paired or unrelated. This flexibility is achieved through the design of segmentation-related text-to-vision alignment, ensuring consistent feature distribution. Additionally, a visual prototype is created to enhance visual representations, which also provides the added benefit of refining the textual features.

This work was supported in part by the National Natural Science Foundation of China under Grant 62171377 and Grant 92470101, in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025019. (Corresponding authors: Y. Xia and Y. Xie.)

Q. Zeng, H. Luo, Z. Lu, Y. Zhang, and Y. Xia are with School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: qjzeng, lhuan, luzl@mail.nwpu.edu.cn; ynzhang, yxia@nwpu.edu.cn)

Y. Xie was with the University of Adelaide, Adelaide, SA 5005, Australia. She is now with the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE (e-mail: yutong.xie678@gmail.com). Her contribution was made when working at the University of Adelaide.

Z. Wang is with the School of Computer Science, The University of Sydney, NSW 2006, Australia (zhiyong.wang@sydney.edu.au).

I. Introduction

ACCURATE medical image segmentation is crucial for effective clinical practice [1]. Recent advances in data-driven deep learning models have significantly enhanced computer-aided diagnosis by improving image interpretation and language processing [2], [3]. Developing high-performance segmentation models often requires careful design of vision models that incorporate elements such as attention mechanisms, diverse backbones, or coarse-to-fine strategies to enhance image processing capabilities [4]. However, the performance of these models is often constrained by the

quantity and quality of available training data.

One promising solution to this limitation is to integrate additional knowledge beyond the images themselves [2], [5]. Textual data, such as clinical reports, are generated alongside medical images in clinical settings. This textual information can offer detailed insights into the location and severity of diseases, thereby providing valuable context for image segmentation tasks [6]. Consequently, there is growing interest in combining text and image data to improve the accuracy and effectiveness of medical image segmentation [7].

Recent developments in large-scale pre-trained vision-language models (VLMs) and language models (LMs) have demonstrated their capacity to understand textual semantic concepts. VLMs, which leverage paired image-text data, establish semantic relationships between visual and textual features [8]. Similarly, LMs, through extensive analysis of vast corpora, acquire a nuanced understanding of context [9], [10]. Both VLMs and LMs offer significant opportunities for effectively combining textual and visual information.

In the field of medical image segmentation, two prominent approaches have emerged for incorporating supplementary textual information into the segmentation process. The first approach [6], [11], [12] employs separately pre-trained vision and text models (see Fig.1, top left). However, this approach faces challenges due to semantic shifts between vision and textual features, as they originate from distinct semantic spaces. While methods like adding learnable prompts and projection layers may help mitigate this issue, no explicit mechanism exists to ensure the alignment of textual and visual features. The second approach [13], [14] employs VLMs, where vision and text models are pre-trained on paired image-text data (see Fig. 1, top right). While this method can initially generate aligned features, the lack of alignment constraints between the visual and textual encoders during segmentation can cause these features to become misaligned over time. Moreover, VLMs typically align features at a high semantic level, which is more suited to recognition tasks but may not fully satisfy the dense prediction requirements of segmentation.

To address these challenges, we propose a novel text-enhanced medical image segmentation method called TeViA, which leverages **Te**xt insights through **Vi**sual **A**lignment. As shown in Fig. 1 (bottom), TeViA ensures the alignment of feature distributions, optimizing the use of textual information in the segmentation process. Unlike conventional methods that rely on adding learnable parameters to implicitly adapt features, TeViA incorporates a regularization term explicitly designed to align textual and visual features. Specifically, for each training sample, a foreground visual feature is generated by multiplying the mask with features extracted from the penultimate layer of the segmentation head. This foreground feature then supervises the adaptation of textual features through projection layers. By enforcing this alignment constraint, TeViA adjusts textual features to a distribution that better supports the segmentation task. This process allows TeViA to effectively integrate textual data for segmentation, with minimal dependence on the specific encoder type used. Additionally, TeViA constructs a visual prototype by aggregating foreground features in a momentum-updated manner. This prototype, which encapsulates target information from all training data, serves as a comprehensive reference for establishing feature-level connections across data instances. While this approach primarily improves visual representations, it also indirectly refines textual features through guidance from the visual features. By employing these techniques, TeViA achieves state-of-the-art performance on benchmark datasets, surpassing a wide range of advanced methods.

The contributions of this work are four-fold:

- We emphasize the importance of feature alignment in integrating text data for medical image segmentation and propose a segmentation-specific text-to-vision (T2V) alignment constraint to adjust textual features.
- We introduce a historical visual prototype that enhances the visual representations of each data instance, while concurrently improving the corresponding textual features.
- We validate the effectiveness and versatility of TeViA through comprehensive experiments, demonstrating that TeViA outperforms the second-best method, LanGuideMedSeg [11], and the leading VLM, CXR-CLIP [13], by 1.58% and 2.29% mIoU scores, respectively, on the QaTa-COV19 dataset, which contains over 2,000 test images.
- We demonstrate the broad applicability of TeViA by consistently improving performance across various vision and text models.

II. Related Work

A. Medical Image Segmentation

Medical image segmentation is widely recognized as a resource-intensive and technically demanding task that often requires expert clinical annotations. With the rise of deep learning, various paradigms such as self-supervised learning [33], [34], few-shot learning [16], [35]–[37], weakly-supervised learning [38]–[40], and semi-supervised learning [18], [41]–[43] have been proposed to reduce reliance on large-scale labeled datasets while maintaining segmentation accuracy. More recently, there has been growing interest in integrating textual knowledge into vision models to further enhance model generalizability and interpretability. In this paper, we summarize and compare representative methods across different learning paradigms and modality settings in Table I, covering both vision-only and vision-language frameworks. This taxonomy highlights the strengths and limitations of each approach and motivates our focus on text-enhanced segmentation methods. Specifically, we aim to explore how textual descriptions, when effectively aligned with visual features, can provide complementary semantic guidance and improve segmentation performance in complex medical scenarios.

B. VLMs and LMs

Recent advancements in VLMs have significantly improved the fusion of textual and visual information. These models excel at combining heterogeneous features, leveraging the complementary strengths of both modalities. A key breakthrough

TABLE I
COMPARISON OF REPRESENTATIVE METHODS ACROSS VARIOUS SUPERVISION SETTINGS AND MODALITY CONFIGURATIONS.

Vision-only				Vision+Language
Self-supervised Learning	**Few-shot Learning**	**Weakly-supervised Learning**	**Semi-supervised Learning**	**Text-enhance Learning**
Swin UNETR [15]: a Swin Transformer-based pre-training framework incorporating proxy tasks such as inpainting, contrastive learning, and rotation prediction. **Pros**: Learns generalized representations by leveraging large-scale unlabeled data. **Cons**: Requires more evidence to support its transferability beyond CT imaging.	**HGRE** [16]: a geometry-driven framework for rare disease scenarios, incorporating uncertainty estimation and adversarial proxy generation. **Pros**: Achieves greater robustness by generating adversarial proxies from limited samples. **Cons**: Incurs additional computational cost due to the construction of a feature memory bank.	**WeakMedSAM** [17]: a SAM-based framework guided by class-level label supervision. **Pros**: Enhances feature representation by leveraging a powerful foundation model. **Cons**: Requires task-specific design to effectively capture fine-grained structures.	**VerSemi** [18]: a unified semi-supervised framework for joint processing of multi-source datasets. **Pros**: Improved generalization and performance across diverse datasets. **Cons**: Increased training time due to multiple forward passes during optimization.	**LViT** [6]: introduces a fine-grained attention module to fuse textual and visual features from parallel branches using diverse pooling strategies. **Pros**: Strengthens visual representation through multi-modal feature fusion. **Cons**: Susceptible to distribution mismatch between textual and visual modalities.
UniMiSS [19]: a universal self-supervised representation learning framework capable of handling both 2D and 3D data. **Pros**: Overcomes the dimensionality gap via a switchable patch embedding module. **Cons**: Suffers from imbalanced data scales between 2D and 3D during pre-training.	**DCGG** [20]: a gradient-guided framework for diagnosing both common and rare diseases, leveraging an optimal transport mechanism. **Pros**: Demonstrates good performance across common and rare categories in few-shot settings. **Cons**: Involves complex knowledge transfer due to channel decomposition.	**Swin-MIL** [21]: a Transformer-based multiple instance learning (MIL) framework. **Pros**: Captures long-range dependencies via the self-attention mechanism. **Cons**: Limited ability to model fine-grained local features due to reduced local receptive field.	**PICK** [22] masks pseudo-label regions to better exploit unlabeled data while reducing error propagation. **Pros**: Mitigates the impact of incorrect predictions on the decoder. **Cons**: Incurred training overhead due to sequential training strategy.	**LanGuideMedSeg** [11]: introduces a GuideDecoder that integrates multi-scale visual and textual features during the decoding process. **Pros**: Enables fine-grained visual decoding with text-guided enhancement. **Cons**: May suffer from suboptimal alignment between visual and textual features.
MSD [23]; CHAOS [24]	**SD-198 [25]; Kvasir [26]**	**BraTS 2019 [27]; H&E [28]**	**LA [29]; NIH-Pancreas [30]**	**MosMedData+ [31]; MoNuSeg [32]**

in this field is CLIP [8], which employs 4 million paired image-text samples for self-supervised contrastive learning and has demonstrated remarkable generalization across a variety of vision tasks. In the medical domain, modality-specific VLMs have been developed for applications such as pathology image analysis [44], [45] and Chest X-Ray diagnosis [13], [14], [46], [47]. In parallel, LMs [9], [48]–[50] have enriched visual tasks by extracting insights from text data. The development of both VLMs and LMs highlights the transformative potential of text in enhancing visual tasks. However, VLMs and LMs serve different roles: VLMs are primarily designed for open-set recognition tasks, where the strong correlation between visual and textual features at a high semantic level is beneficial [44], [51], while LMs are more flexible and can be seamlessly integrated as supplementary branches in vision tasks. In this study, we explore the use of textual data to enhance medical image segmentation, where both VLMs and LMs contribute to knowledge transfer between modalities.

C. Text-enhanced Medical Image Segmentation

With the development of VLMs and LMs [5], [52], [53], there is a growing trend of incorporating textual information into medical image segmentation. Several approaches have been proposed to leverage textual data [54], [55]. For instance, the CLIP-driven Universal Model [56] uses the CLIP text encoder to assess relationships among organs, dynamically incorporating textual knowledge into the segmentation head. Similarly, LViT [6] uses a pre-trained BERT model [57] to extract information from medical reports, improving segmentation performance, particularly for lesion regions. LanGuideMedSeg [11] utilizes CXR-BERT [58], a specialized LM for Chest X-Ray images, to improve segmentation accuracy. TPRO [40] incorporates ClinicalBERT [59] for histopathology tissue segmentation, enriching the process with detailed semantic information from text descriptions.

Despite these advancements, current text-enhanced methods typically rely on learnable prompts and layers to combine textual and visual features directly, aiming for adaptive integration of knowledge. However, since textual and visual features stem from different models pre-trained for distinct objectives, a semantic shift naturally occurs between them. This shift can limit the ability of textual features to contribute effectively to vision tasks, despite the potential information they offer. In this paper, we advocate for aligning textual cues with foreground visual features, thereby ensuring better integration of textual knowledge into the segmentation process and addressing the semantic discrepancy between modalities. This approach aims to enhance overall segmentation performance by effectively utilizing textual information.

III. METHOD

A. Preliminaries

In this study, the dataset is structured as $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$, where x_i denotes the i-th image, t_i represents the corresponding text description, y_i is the associated mask, and N indicates the number of data points. Our objective is to train a segmentation model $f(\cdot;\theta)$ such that $f(x_i, t_i; \theta) \rightarrow y_i$. Specifically, $f(\cdot;\theta)$ consists of a vision encoder $f(\cdot;\theta_v)$, a text encoder $f(\cdot;\theta_t)$ and a decoder $f(\cdot;\theta_d)$ that integrates textual and visual features to produce the final segmentation map.

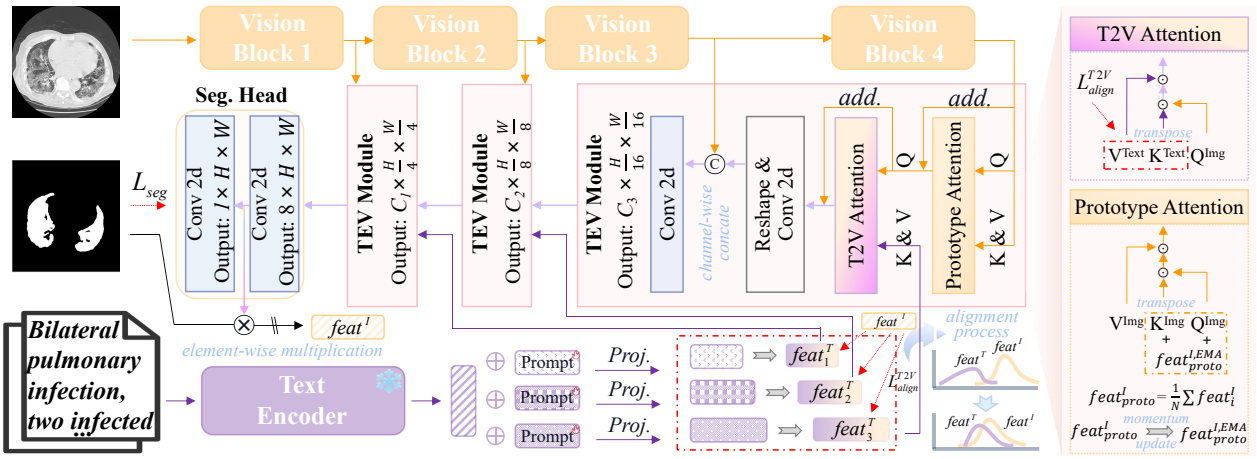


Fig. 2. Illustration of our proposed TeViA. The upper branch is the trainable vision encoder, divided into four blocks. The bottom branch represents the projection of textual features using a frozen text encoder guided by learnable prompts that are randomly initialized and optimized during training to better adapt textual knowledge to the segmentation task. Specifically, the foreground visual feature is utilized as the supervision signal to help textual features shift towards a distribution that benefits the segmentation. The middle branch depicts the decoding process. Here, text-to-vision (T2V) attention is employed to fuse textual and visual features, while prototype attention is designed to emphasize significant features of each data point using a momentum-aggregated historical visual prototype. The decoding process sequentially applies prototype attention and T2V attention, both implemented using a multi-head self-attention mechanism. The right side provides details of the two attention modules.

B. Overview of TeViA Framework

The proposed TeViA framework features three primary branches. The vision encoding branch $f(\cdot;\theta_v)$ handles visual features. The text encoding branch $f(\cdot;\theta_t)$ manages textual features. To ensure the effective integration of textual features into the segmentation task, a Text-to-Vision (T2V) alignment is performed. The last branch $f(\cdot;\theta_d)$ is the decoding process, which gradually integrates textual and visual features from each stage using the Text-enhanced Vision Module (TEV Module) with prototype attention and T2V attention. Note that the architectures of $f(\cdot;\theta_v)$ and $f(\cdot;\theta_t)$ are flexible and can be adapted to kinds of backbones. The pipeline of our TeViA framework was illustrated in Figure 2. Now we delve into details.

C. TEV Module

Given an image x_i with a text description t_i, the visual feature at stage s is computed as:

$$feat^I_{i,s} = f(x_i;\theta^s_v). \tag{1}$$

The corresponding textual feature is first generated by $f(\cdot;\theta_t)$, and then produced by an MLP layer with a learnable prompt $[Prompt_{\#s}]$, shown as follows:

$$feat^T_{i,s} = MLP(f(t_i;\theta_t) + [Prompt_{\#s}]). \tag{2}$$

Note that the MLP layer ensures the visual and textual features have the same dimension.

Prototype Attention. Unlike traditional self-attention, we introduce a visual prototype that captures the foreground semantics across all training data to enhance each data point's visual representation. The foreground visual feature is calculated as:

$$feat^I_{i,fg} = feat^I_i \odot y_i, \tag{3}$$

where $feat^I_i$ is the visual feature from the penultimate layer of the segmentation head, and $\odot$ denotes element-wise multiplication. Note that $feat^I_i$ has the same spatial shape as the mask y_i. The current visual prototype is obtained by averaging the foreground features:

$$feat^I_{proto} = \frac{1}{N}\sum_{i=1}^{N} feat^I_{i,fg}, \tag{4}$$

The visual prototype is then momentum-updated as:

$$feat^{I,EMA}_{proto} = \alpha feat^{I,EMA}_{proto} + (1-\alpha) feat^I_{proto}, \tag{5}$$

where the momentum factor α is set to 0.99. To match feature dimensions across stages, $feat^{I,EMA}_{proto}$ is interpolated to $feat^{I,EMA}_{proto,s}$. The enhanced visual representation of image x_i at scale s is derived from:

$$feat^{I,P}_{i,s} = Attn^{Proto}(P,Q,K,V), \tag{6}$$

where $Attn^{Proto}$ performs prototype attention:

$$Attn^{Proto}(P,Q,K,V) = softmax(\frac{(Q+P)(K+P)^T}{\sqrt{d_k}})V, \tag{7}$$

where d_k is the dimension of the feature. The prototype P is represented by $feat^{I,EMA}_{proto,s}$, and $feat^I_{i,s}$ is used as the query Q, key K, and value V. The refined visual feature $feat^{I,P}_{i,s}$ is then ready for integration with the textual feature.

T2V Attention. This module fuses the textual feature $feat^T_{i,s}$ with the enhanced visual feature $feat^{I,P}_{i,s}$. This process is written as:

$$feat^{I,T}_{i,s} = Attn^{T2V}(Q,K,V), \tag{8}$$

where $feat^{I,T}_{i,s}$ is the fused feature at scale s, and $Attn^{T2V}$ is the T2V attention that can be computed as:

$$Attn^{T2V}(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{9}$$

where $Attn^{T2V}$ follows the standard cross-attention calculations, with $feat^{I,P}_{i,s}$ serving as the query Q, and $feat^T_{i,s}$ acting

as both the key K and value V. After fusion, $feat_{i,s}^{I,T}$ is upsampled using a convolutional layer to match the feature size at scale $s-1$, and is concatenated with $feat_{i,s-1}^{I}$ channel-wise in a resudual manner, generating the refined feature at scale $s-1$. So far, the TEV module has successfully fused the visual and textual features at scale s, and provided features for scale $s-1$.

Each decoding stage follows a consistent process: first, it refines the visual features, and then it integrates the textual features. The final prediction is generated by a segmentation head, which consists of two convolutional layers.

D. Objective Function and Inference

Objective Function. TeViA is optimized with two loss functions: a segmentation loss and a proposed text-to-vision (T2V) alignment loss. The segmentation loss, denoted as $\mathcal{L}_{seg}$, includes both Dice and Cross-Entropy terms:

$$\mathcal{L}_{seg} = L_{\text{Dice}}(f(x_i, t_i; \theta), y_i) + L_{\text{CE}}(f(x_i, t_i; \theta), y_i), \quad (10)$$

which guides the model to produce both region-consistent (Dice) and voxel-wise accurate (CE) predictions.

T2V Alignment Loss. To ensure that textual semantics are meaningfully aligned with the task-relevant visual regions—rather than merely fused via standard attention—we introduce the T2V alignment loss $\mathcal{L}_{align}$ as a core component of TeViA. This loss explicitly regularizes the projected textual features to align with foreground visual features, thereby enforcing segmentation-aware cross-modal consistency. During training, the foreground visual feature $feat_{i,\text{fg}}^{I}$ is extracted by applying the predicted segmentation mask to the visual encoder's intermediate representations. These features are then interpolated to match the spatial dimensions of the multi-scale textual features $feat_{i,s}^{T}$, resulting in $feat_{i,s,\text{fg}}^{I}$ for alignment at each stage s. The T2V alignment loss is defined as:

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{3}\beta_s\left(1 - \frac{feat_{i,s}^{T}\cdot feat_{i,s,\text{fg}}^{I}}{\|feat_{i,s}^{T}\|_2 \cdot \|feat_{i,s,\text{fg}}^{I}\|_2}\right), \quad (11)$$

where $feat_{i,s}^{T}$ denotes the textual features projected to stage s, and cosine similarity is used to measure alignment. The hyperparameters $\{\beta_1, \beta_2, \beta_3\}$ balance the contributions of different feature scales. This design serves three purposes: (1) it encourages the textual features to align with visual regions actually contributing to segmentation predictions, (2) it avoids degenerate fusion where textual features are underutilized or misaligned, and (3) it provides a *plug-and-play constraint* that is model-agnostic, making it compatible with various pretrained text encoders, vision encoders, and vision-language models. To this end, the total training objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda\mathcal{L}_{align}, \quad (12)$$

where λ is a hyperparameter that balances the two loss terms.

Inference. For an unseen image with a corresponding text description, such as *"Bilateral pulmonary infection, two infected areas, all left lung and middle right lung"*, the vision branch $f(\cdot;\theta_v)$ and text branch $f(\cdot;\theta_t)$ generate visual and textual features, respectively. The decoder $f(\cdot;\theta_d)$ then fuses and refines these features to produce the final results.

IV. Experiments

A. Setup

Datasets. We utilized five publicly available datasets for evaluation. The MosMedData+ dataset [31] comprises 2,729 CT scan slices targeting lung infection segmentation, while the QaTa-COV19 dataset [60] contains 9,258 COVID-19 Chest X-Ray images for the same task. Both datasets include associated textual descriptions detailing the number and location of infected areas, provided by [6]. For fair comparison with existing methods, data splits adhered to those established in [6] and [12]: MosMedData+ was divided into 2,183 training, 273 validation, and 273 test images, and QaTa-COV19 was split into 5,716 training, 1,429 validation, and 2,113 test images. The MoNuSeg dataset [32] consists of 30 digital microscopic tissue images, focusing on nuclear segmentation. Given its limited size, we employed three-fold cross-validation for evaluation. The Breast Tumor dataset [61], [62] contains 763 ultrasound images for breast tumor boundary delineation, accompanied by textual descriptions detailing tumor shape, echogenicity, and margin characteristics. Finally, the Brain Tumor dataset [63] includes 3,064 MRI scans, with text-based annotations describing tumor regions for brain tumor segmentation. The text prompts for both the Breast Tumor and Brain Tumor datasets were sourced from [63]. Collectively, these datasets enable a comprehensive assessment of our method's performance across diverse imaging modalities (CT, X-ray, pathology, ultrasound, and MRI) and varied segmentation tasks (lung infection, cell, breast tumor, and brain tumor segmentation).

Implementation Details. Model training was conducted using the AdamW optimizer with a batch size of 32. Input images were resized to 224×224 pixels, and data augmentation techniques, including random cropping and random zooming, were applied. We implemented a cosine annealing learning rate schedule, starting from an initial rate of 3e-4 and decreasing to a minimum of 1e-6. All experiments were performed using PyTorch [64] on a single NVIDIA Tesla P100 GPU with 16GB VRAM. Following [6], [12], we evaluated model performance using Dice coefficient and mean Intersection over Union (mIoU) metrics.

B. Results on the MosMedData+ and QaTa-COV19

We compared our approach with both vision-only and vision+language methods. Specifically, Table II presents the model performance on the MosMedData+ and QaTa-COV19 datasets. The detailed analysis is outlined below.

Results of vision-only methods. As shown in Table II, we evaluated six prominent vision-only methods, including U-Net [65], U-Net++ [66], nnUNet [67], TransUNet [68], Swin-Unet [69], UCTransNet [70] and ACC-UNet [71]. These models incorporate various vision backbones, such as ConvNeT [72], ViT [51], Swin-T [73], and ConvNeXt [74]. Among them, ACC-UNet, which uses ConvNeXt, achieved the highest accuracy. Our TeViA approach surpasses ACC-UNet by 5.87% and 8.15% in Dice scores across both datasets, demonstrating the efficacy of integrating textual knowledge.

TABLE II

PERFORMANCE COMPARISONS ON THE MOSMEDDATA+ AND QATA-COV19 DATASTES. [VLM] REFERS TO METHODS THAT EMPLOY PAIRED VISION AND TEXT ENCODERS FROM PRE-TRAINED VLMS, WHILE [V+T] DENOTES METHODS THAT UTILIZE INDEPENDENT VISION AND TEXT ENCODERS. THE SYMBOL † INDICATES THAT THE IMPLEMENTATION AND RESULTS ARE SOURCED FROM [6]. SYMBOL ∗ DENOTES P-VALUE < 0.05 WITH THREE REPEATED EXPERIMENTS.

Methods	V-Encoder	T-Encoder	MosMedData+		QaTa-COV19		Param (M)	Flops (G)
			Dice (%)	mIoU (%)	Dice (%)	mIoU (%)		
Vision only								
U-Net	Conv	N/A	64.60	50.73	79.02	69.46	14.8	50.3
U-Net++	Conv	N/A	71.75	58.39	79.62	70.25	74.5	94.6
nnUNet	Conv	N/A	72.59	60.36	80.42	70.81	19.1	412.7
Swin-Unet	Swin-T	N/A	63.29	50.19	78.07	68.34	82.3	67.3
TransUNet	Conv+ViT	N/A	71.24	58.44	78.63	69.13	105	56.7
UCTransNet	Conv+ViT	N/A	65.90	52.69	79.15	69.60	65.6	63.2
ACC-UNet	ConvNeXt	N/A	72.64	59.54	82.88	72.74	16.8	45.3
Vision + Language	***init. from VLMs***							
†LViT [VLM]	ConVIRT		72.06	59.73	79.72	70.58	35.2	44.6
†LViT [VLM]	GLoRIA		72.42	60.18	79.94	70.68	45.6	60.8
LanGuideMedSeg [VLM]	MedKLIP		74.45	59.30	88.42	80.86	107.7	22.6
LanGuideMedSeg [VLM]	CXR-CLIP		75.48	62.62	89.53	81.31	161.0	30.6
LAVT [V+T]	Swin-T	BERT	73.29	60.41	79.28	69.89	118.6	83.8
LViT [V+T]	Conv+ViT	BERT	74.57	61.33	83.66	75.11	29.7	54.1
CPAM [V+T]	VGG16	CLIP^{Text}	76.88	62.44	87.53	77.82	58.7	32.2
LanGuideMedSeg [V+T]	ConvNeXt	CXR-BERT	$76.98_{\pm 0.15}$*	$62.74_{\pm 0.33}$*	$90.27_{\pm 0.42}$*	$82.02_{\pm 0.71}$*	146.8	11.2
TeViA (Ours)	ConvNeXt	CXR-BERT	$78.49_{\pm 0.06}$	$64.59_{\pm 0.05}$	$91.06_{\pm 0.09}$	$83.60_{\pm 0.14}$	146.8	11.2

TABLE III

THREE-FOLD CROSS VALIDATION ON THE MONUSEG DATASET.

Methods	Dice ↑		mIoU ↑	
	mean (%)	std (%)	mean (%)	std (%)
U-Net	75.77	0.17	60.99	0.28
nnUNet	77.20	0.07	63.30	0.13
Swin-Unet	75.36	0.48	60.47	0.80
TransUnet	76.53	0.12	61.99	0.20
UCTransNet	78.57	0.24	64.71	0.45
ACC-UNet	78.08	0.12	64.03	0.16
MedKLIP	77.36	0.19	63.08	0.35
CXR-CLIP	79.28	0.34	65.67	0.65
LViT	78.84	0.06	65.07	0.12
CPAM	79.19	0.38	65.94	0.73
LanGuidMedSeg	79.69	0.09	66.24	0.17
TeViA (Ours)	80.49	0.09	67.36	0.18

TABLE IV

RESULTS ON THE BREAST TUMOR AND BRAIN TUMOR DATASET. SYMBOL ∗ DENOTES P-VALUE < 0.05 WITH THREE REPEATED EXPERIMENTS.

Methods	Breast Tumor		Brain Tumor	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
U-Net	77.37	63.09	79.90	66.52
nnUNet	79.14	71.24	81.38	68.51
ACC-UNet	77.65	63.46	78.91	65.16
MedKLIP	75.27	60.35	80.81	67.80
CXR-CLIP	77.42	63.16	76.03	61.33
LViT	65.54	46.92	77.56	63.34
CPAM	69.72	53.52	77.60	63.40
LanGuidMedSeg	$79.82_{\pm 1.22}$*	$66.42_{\pm 1.66}$*	$82.01_{\pm 1.03}$*	$69.55_{\pm 1.49}$*
TeViA (Ours)	$85.71_{\pm 0.37}$	$75.00_{\pm 0.57}$	$83.36_{\pm 0.05}$	$71.46_{\pm 0.07}$

Results of vision + language methods. As show in Table II, we adapted four advanced VLMs for comparison, including ConVIRT [75], GLoRIA [76], MedKLIP [14], and CXR-CLIP [13]. The implementations and results for ConVIRT and GLoRIA were reported by [6], while MedKLIP and CXR-CLIP were integrated into the LanGuideMedSeg architecture [11] by substituting its vision and text encoders with corresponding pre-trained VLMs. It is noteworthy that the vision models within these VLMs were pre-trained on a vast number of external Chest X-Ray images. The results indicate that VLMs generally outperform vision-only methods. For example, LanGuideMedSeg initialized with VLM-MedKLIP achieved 5.54% and 8.12% higher Dice and mIoU scores, respectively, compared to the best vision-only method, ACC-UNet, on the QaTa-COV19 dataset. This indicates the significance of incorporating textual knowledge into segmentation tasks. Compared to LanGuideMedSeg with the leading VLM-CXR-CLIP initialization, our TeViA demonstrates substantial improvements of 3.01% and 1.53% in the Dice metric on the MosMedData+ and QaTa-COV19 datasets, respectively.

We also evaluated four methods that use independent vision and text models for text-enhanced medical image segmentation, including LAVT [77], LViT [6], CPAM [12], and LanGuideMedSeg [11]. Each method employs unique decoder designs for feature integration. Kindly note that the costs of TeViA and LanGuideMedSeg differ slightly. TeViA computes foreground visual features to generate a prototype for alignment, incurring an additional 0.0238M parameters and 0.002G FLOPs. However, due to rounding, the reported values may appear identical. Below are the summarized findings: (1) These methods tend to outperform VLM-initialized ap-

TABLE V

ABLATION STUDIES IN THE FIRST REPETITION OF THE EXPERIMENT. "IMAGE" MEANS USING ONLY IMAGE DATA. "T-ENCODER+ATTN^{T2V}" REPRESENTS EMPLOYING PRE-TRAINED TEXT ENCODER TO ENCODE TEXT DATA, ALONG WITH T2V ATTENTION TO FUSE TEXTUAL AND VISUAL FEATURES. "$\mathcal{L}_{align}^{T2V}$" STANDS FOR T2V ALIGNMENT. "ATTN^{Proto}" IS THE PROTOTYPE ATTENTION WHICH INCORPORATES A MOMENTUM-UPDATED VISUAL PROTOTYPE.

Mark	Image	T-Encoder +Attn^{T2V}	$\mathcal{L}_{align}^{T2V}$	Attn^{Proto}	MosMedData+		QaTa-COV19		Param (M)	Flops (G)
					Dice (%)	mIoU (%)	Dice (%)	mIoU (%)		
Row 1	✓	-	-	-	71.58	58.72	81.36	71.68	34.90	8.53
Row 2	✓	-	-	✓	76.11	61.43	87.65	78.02	36.70	8.54
Row 3	✓	✓	-	-	76.12	61.45	87.57	77.89	145.0	11.1
Row 4	✓	✓	-	✓	77.67	63.45	90.12	82.54	146.8	11.2
Row 5	✓	✓	✓	-	78.09	64.05	90.43	82.75	145.0	11.1
TeViA	✓	✓	✓	✓	**78.51**	**64.62**	**91.03**	**83.54**	146.8	11.2

proaches on datasets with less training samples. For instance, LanGuideMedSeg outperforms the VLM initializations of CXR-CLIP and MedKLIP by 1.41% and 2.44% in Dice scores on the MosMedData+ dataset. This performance difference may be attributed to the varied vision backbones used. ACC-UNet, a ConvNeXt-based model without textual knowledge, achieves comparable results to VLMs like ConVIRT and GLoRIA, suggesting that ConvNeXt may be particularly suited to this dataset. However, due to the availability of pre-trained weights from MedKLIP and CXR-CLIP, ResNet-50 and Swin-T were ultimately used in our experiments. (2) For the QaTa-COV19 dataset, which has a larger training sample size, the performance of these methods is more comparable. For example, the Dice score difference between LanGuideMedSeg and CXR-CLIP is less than 0.7%, indicating that effective use of VLMs requires a substantial amount of paired data. (3) Notably, TeViA shows improvements of 1.85% and 1.58% in mIoU compared to LanGuideMedSeg on the MosMedData+ and QaTa-COV19 datasets, respectively, by consistently utilizing the same vision and text models. This highlights the effectiveness of our proposed design in leveraging textual information for enhanced segmentation performance.

C. Results on the MoNuSeg Dataset

Moreover, to assess TeViA's generalizability across diverse imaging modalities, we evaluated its performance on the histopathological MoNuSeg dataset. As detailed in Table III, TeViA achieved the highest segmentation performance with a Dice score of 80.49% and an mIoU of 67.36%, accompanied by minimal standard deviations (0.09% and 0.18%, respectively). These results not only underscore high segmentation accuracy but also remarkable consistency across diverse samples. Notably, TeViA consistently outperformed existing text-enhanced methods, such as LanGuideMedSeg (79.69% Dice, 66.24% mIoU) and its variant initialized with VLM-CXR-CLIP (79.28% Dice, 65.67% mIoU), across both metrics. This robust performance highlights the adaptability of our approach in handling pathological images, a domain where significant challenges arise from intricate tissue structures and staining variability.

D. Results on the Breast and Brain Tumor Datasets

To further assess its generalizability, TeViA was evaluated on two distinct imaging modalities: ultrasound (breast tumor dataset) and MRI (brain tumor dataset). As detailed in Table IV, TeViA consistently outperforms all competing methods on both datasets. On the breast tumor dataset, TeViA achieves 85.71% Dice and 75.00% mIoU, surpassing the strongest baseline, LanGuidMedSeg (79.82% Dice, 66.42% mIoU), by 5.89% Dice and 8.58% mIoU. Compared to the classic U-Net (77.37% Dice, 63.09% mIoU), the gains are even more substantial, reaching 8.34% Dice and 11.91% mIoU. Similarly, on the brain tumor dataset, TeViA attains the highest performance with 83.36% Dice and 71.46% mIoU. It outperforms LanGuidMedSeg by 1.35% Dice and 1.91% mIoU, and exceeds ACC-UNet by 4.45% Dice and 6.30% mIoU. These consistent improvements highlight TeViA's effectiveness in leveraging lightweight textual prompts to enhance segmentation accuracy, even under challenging conditions like low-contrast ultrasound and heterogeneous brain MRI. Moreover, by demonstrating superior alignment between visual and textual cues when compared to existing vision-only and vision-language baselines (e.g., MedKLIP and CXR-CLIP), TeViA reinforces its versatility and robust domain-adaptability for multimodal clinical environments.

E. Ablation Study

Table V illustrates the impact of each module within TeViA in the first repetition of the experiment. The following detailed analysis provides insights into these effects.

Effect of Text Knowledge. Comparison of Row 3 with Row 1 demonstrates the significance of integrating text data for enhancing segmentation. When using ConvNeXt-Tiny as the vision encoder, incorporating CXR-BERT and T2V attention for extracting and combining textual features results in approximately a 5% improvement in Dice scores across both datasets. This underscores the effectiveness of leveraging supplementary textual knowledge to improve segmentation performance.

Effect of $\mathcal{L}_{align}^{T2V}$. The comparison between Row 5 and Row 3 highlights the effectiveness of the T2V alignment $\mathcal{L}_{align}^{T2V}$. With this alignment constraint in place, TeViA outperforms all competitors. Specifically, TeViA achieves Dice scores of 78.09%

TABLE VI
PLUG-AND-PLAY EFFECT OF THE TEXT-TO-VISION ALIGNMENT $\mathcal{L}_{align}^{T2V}$, WHICH BOOSTS OTHER METHODS SIGNIFICANTLY.

Method	$\mathcal{L}_{align}^{T2V}$	MosMedData+		p-value
		Dice (%)	mIoU (%)	
MobileNet (LanGuide)	w/o	70.51	54.45	<0.01
	w	73.95 3.44 ↑	58.68 4.23 ↑	
U-Net (LanGuide)	w/o	74.60	59.49	<0.01
	w	76.56 1.96 ↑	62.01 2.52 ↑	
GLoRIA	w/o	72.42	60.18	<0.01
	w	77.81 5.39 ↑	63.68 3.50 ↑	
LViT	w/o	74.57	61.33	<0.01
	w	77.92 3.35 ↑	63.89 2.56 ↑	
MedKLIP	w/o	74.45	59.30	<0.01
	w	78.07 3.62 ↑	64.04 4.74 ↑	
CXR-CLIP	w/o	75.48	62.62	<0.01
	w	78.25 2.77 ↑	64.27 1.65 ↑	
LanGuideMedSeg	w/o	77.10	62.75	<0.05
	w	78.09 0.99 ↑	64.05 1.30 ↑	

and 90.43% on the respective datasets, showing improvements of 0.99% and 0.65% compared to the second best method, LanGuideMedSeg. These results indicate that aligning textual features with the corresponding foreground visual features is crucial for effective segmentation.

Effect of Visual Prototype Reference. The visual prototype is a momentum-building feature that aggregates the foreground semantics from all training data. Its impact is evident when comparing Row 2 to Row 1, with significant Dice improvements of 4.53% and 6.29% observed on the MosMedData+ and QaTa-COV19 datasets, respectively. These improvements highlight the benefit of the visual prototype in enhancing visual representations by incorporating historical foreground semantics into each data instance.

Improved Visual Representations Enhance Textual Representations. Row 4 shows higher results compared to Row 3, indicating that the introduction of Attn^{Proto} improves visual representations beyond the benefits provided by textual knowledge alone. Furthermore, comparing the final row with Row 4 reveals additional improvements, attributed solely to the $\mathcal{L}_{align}^{T2V}$ constraint, which regularizes textual features using visual ones. This demonstrates that refined visual representations can enhance textual representations correspondingly, validating the reciprocal relationship between visual and textual features.

V. DISCUSSION

A. Plug-and-play Effect

The primary feature of TeViA is the $\mathcal{L}_{align}^{T2V}$ constraint, which can be seamlessly integrated into other methods with minimal additional computational cost. We explored the effects of $\mathcal{L}_{align}^{T2V}$ on various competing methods. All evaluated VLMs were adapted by replacing their respective vision and text encoders within the LanGuideMedSeg architecture. Crucially, the decoder's structure remained consistent across all variants: it first employs a self-attention module for visual representation refinement, followed by a cross-attention module that fuses visual and textual features. Our alignment loss $\mathcal{L}_{align}^{T2V}$ is applied within this cross-attention module, where textual features are explicitly aligned with aggregated foreground visual representations, mirroring the strategic alignment used in our TeViA framework. As evidenced in Table VI, the incorporation of $\mathcal{L}_{align}^{T2V}$ yielded significant improvements across all methods, with a p-value < 0.05. For instance, $\mathcal{L}_{align}^{T2V}$ resulted in a notable 3.62% Dice score improvement for MedKLIP initialization and a 1.30% increase in mIoU for the second-best method, LanGuideMedSeg. A larger performance gain of 3.44% in Dice was observed when applying $\mathcal{L}_{align}^{T2V}$ to the lightweight MobileNet encoder [78] within the LanGuideMedSeg architecture. These results highlight the versatile and broad applicability of our alignment strategy. In addition, $\mathcal{L}_{align}^{T2V}$ yields more notable improvements when applied to other models compared to its integration within the proposed TeViA framework. This is primarily because TeViA is built upon a stronger baseline, with visual representations already deeply refined through semantic foreground prototypes.

TABLE VII
DISCUSSION ON THE IMPACT OF TEXT ENCODERS ON TEVIA'S RESULTS, WITH CONVNEXT-TINY AS THE VISION MODEL.

T-Encoder	MosMedData+		QaTa-COV19	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
CLIP^{Text}	77.75	63.60	90.63	82.86
CXR-CLIP^{Text}	78.04	63.98	90.79	83.14
BioClinicalBERT	78.30	64.34	90.49	82.63
RadBERT	78.42	64.50	90.89	83.31
CXR-BERT	**78.51**	**64.62**	**91.03**	**83.54**

B. Flexibility with Diverse Text Models

We further investigated the effect of different text models on TeViA, using ConvNeXt-Tiny as the vision backbone. This included LMs such as BioClinicalBERT [59], RadBERT [48] and CXR-BERT [58], as well as text encoders from pre-trained VLMs like CLIP and CXR-CLIP. Our findings, summarized in Table VII, reveal the following: (1) TeViA's performance remains consistent across different text encoders, with only minor variations. This stability is largely attributed to the T2V alignment, which effectively regularizes textual representations to align with foreground visual features. (2) The first row of Table VII shows TeViA using the CLIP text encoder, which is pre-trained on internet data without specific medical knowledge. Notably, even under this condition, TeViA performs comparably, with slight improvements over the second-best LanGuideMedSeg, which uses the medical domain-specific CXR-BERT pre-trained on Chest X-Ray reports. This observation suggests that an effective alignment strategy may be more crucial than the choice of pre-trained model when dealing with heterogeneous data.

C. Collaborating with Diverse Vision Backbone

Table VIII shows the performance of TeViA with various vision models, including Vit-Tiny, Swin-Tiny, ResNet-50 and ConvNeXt-Tiny, all utilizing pre-trained weights from ImageNet [79]. The results indicate that (1) the choice of vision

TABLE VIII
FLEXIBILITY OF TEVIA WITH VARIOUS VISION BACKBONES, USING CXR-BERT AS THE TEXT ENCODER.

V-Encoder	MosMedData+		QaTa-COV19	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
ViT-Tiny	77.13	62.86	90.00	81.82
Swin-Tiny	76.83	62.37	89.95	81.73
ResNet-50	77.25	63.09	90.47	82.60
ConvNeXt-Tiny	**78.51**	**64.62**	**91.03**	**83.54**

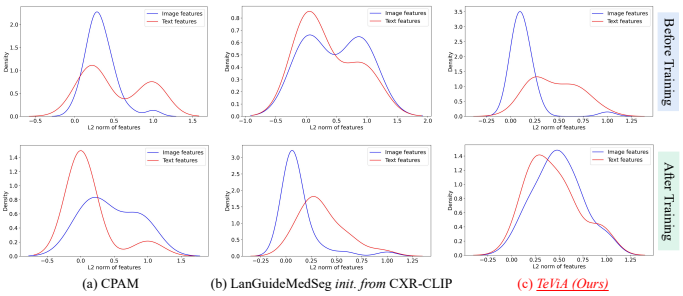
TABLE IX
QUANTITATIVE ANALYSIS OF FEATURE ALIGNMENT ON THE QATA-COV19 TEST SET.

	CPAM	LanGuideMedSeg	TeViA (Ours)
Cosine Similarity	- 0.1933	0.1147	0.4612

model has a more significant impact on TeViA's performance than the choice of text model. For example, on the MosMedData+ dataset, changing vision models causes a Dice score fluctuation of 1.68%, compared to only 0.76% with different text models. This suggests that the vision model plays a more critical role in segmentation performance. We attribute this phenomenon to two main reasons. First, the prediction relies primarily on the input image with text data serving as supplementary information. Consequently, the vision model is more crucial than the text model in producing high-quality representations. Second, since textual features are aligned with visual ones, the quality of textual knowledge extraction is closely linked to the quality of visual representations; thus, poorer visual representations lead to less effective utilization of text data. (2) Despite these variations in performance, TeViA outperforms LanGuideMedSeg with CXR-CLIP initialization by 1.35% in Dice score on the MosMedData+ dataset when using the Swin-Tiny model. It also surpasses MedKLIP initialization by 1.74% in mIoU on the QaTa-COV19 dataset with the same ResNet-50 model. These results, achieved without relying on additional medical knowledge from pre-training data, further validate the effectiveness of TeViA in improving visual representations and leveraging textual info.

D. Kernel Density Estimation

Fig. 3 presents the kernel density estimation (KDE) of textual and visual features, both before and after training on the QaTa-COV19 dataset. KDE was employed to visualize the distributional similarity between these features once projected into a shared latent space. A greater overlap between the two distributions after training indicates improved alignment between the modalities. Our findings are as follows. (1) Methods initialized from pre-trained VLMs, such as LanGuideMedSeg [11] (which leverages CXR-CLIP [13]), inherently exhibit a preliminary alignment between textual and visual features prior to fine-tuning. (2) Without constraints between visual and textual features during segmentation training, models like CPAM [12] and LanGuideMedSeg mismatched feature distributions. This suggests that achieving adaptive feature integration remains challenging, even with VLM initialization.



(a) CPAM (b) LanGuideMedSeg *init. from* CXR-CLIP (c) *TeViA (Ours)*

Fig. 3. Kernel density estimation on the test set of QaTa-COV19 dataset. The KDE was performed on the L2 norms of the feature vectors extracted from the image and text encoders. A strong alignment between the modalities would result in similar or overlapping density distributions, indicating that the visual and textual features are semantically close and follow a similar pattern across the dataset. The top row shows the distribution of visual and textual features before training, while the bottom row displays the distribution after training. CPAM [12] and LanGuideMedSeg [11] initialized with CXR-CLIP [13] are listed for comparison.

TABLE X
RESULTS IN THE ANNOTATION-EFFICIENT SETTING. COMPARISONS ARE MADE WITH SEMI-SUPERVISED METHODS, INCLUDING CPS [80], BCP [81], AND LEFED [82].

Method	Labeled	Unlabeled	Text	MosMedData+	
				Dice (%)	mIoU (%)
CPS	10%	90%	×	65.50	49.63
BCP	10%	90%	×	66.43	50.78
LeFeD	10%	90%	×	67.75	51.32
TeViA	10%	×	✓	69.97	53.81
CPS	20%	80%	×	67.87	51.52
BCP	20%	80%	×	69.21	53.30
LeFeD	20%	80%	×	70.13	54.09
TeViA	20%	×	✓	71.21	55.29
CPS	50%	50%	×	70.72	55.18
BCP	50%	50%	×	71.38	55.76
LeFeD	50%	50%	×	72.06	56.21
TeViA	50%	×	✓	74.00	58.74

(3) In contrast, our TeViA framework demonstrates well-aligned features, underscoring the effectiveness of aligning textual features with the foreground visual ones via our T2V alignment loss. Furthermore, a quantitative evaluation presented in Table IX shows that TeViA achieves a significantly higher cosine similarity score (0.4612) compared to LanGuideMedSeg (0.1147) and CPAM (-0.1933), highlighting TeViA's superior ability to bridge the semantic gap between visual and textual representations.

E. Pure Image data vs Image-text Data: A Case Study under Annotation-efficient Scenario

Table X showcases the results of TeViA using 10%, 20% and 50% image-text data, compared with semi-supervised methods CPS [80], BCP [81], and LeFeD [82], which use 10%, 20%, and 50% image annotations. The results indicate the following: (1) TeViA consistently achieves the highest performance across all data utilization scenarios, underscoring the significance of incorporating supplementary text data alongside images. (2) With 10% image-text data, TeViA

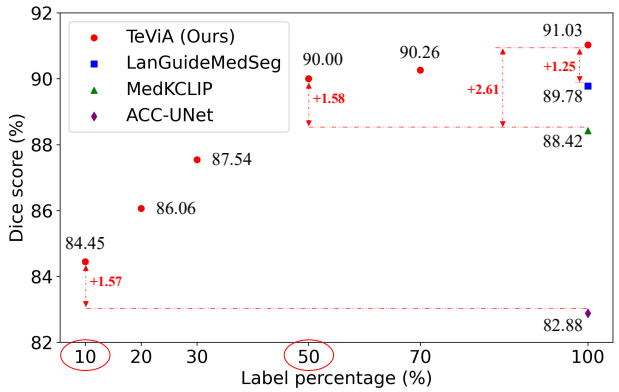


Fig. 4. TeViA's performance under label percentages of 10%, 20%, 30%, 50%, 70% and 100%, on the QaTa-COV19 dataset. It is surprising to find that TeViA is able to surpass all competitors with 50% annotations.

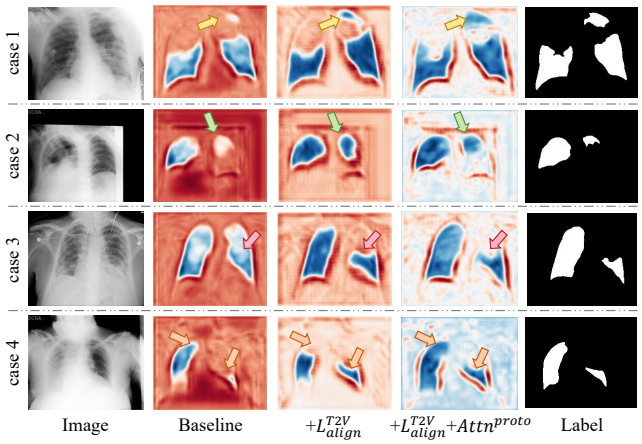


Fig. 5. Feature visualization. These features are derived from the penultimate layer of the segmentation head. The different regions are marked by arrow.

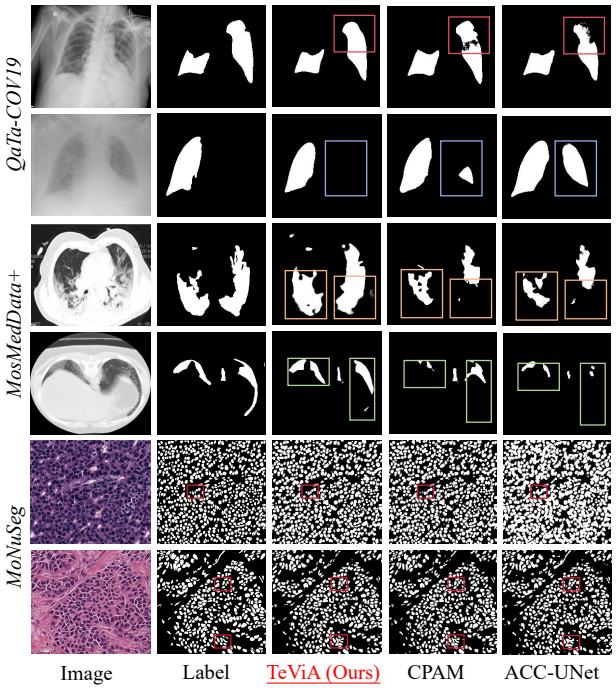


Fig. 6. Segmentation results. The top two cases are collected from the QaTa-COV19 dataset, the middle two cases are from the MosMedData+ dataset, and the bottom two cases are derived from the MoNuSeg dataset. The different regions are highlighted by boxes.

TABLE XI
THE IMPACT OF VARIOUS ALIGNMENT LOSS FUNCTIONS ON THE TEXT-TO-VISION ALIGNMENT PROCESS.

Loss Function	QaTa-COV19	
	Dice (%)	mIoU (%)
InfoNCE	90.84	83.22
KL Divergence	90.95	83.41
Wasserstein Distance	90.99	83.46
Cosine Similarity	**91.03**	**83.54**

demonstrates improvements of 2.22% in Dice and 2.49% in mIoU compared to LeFeD, which relies on 10% labeled images and 90% unlabeled images. These substantial gains further highlight TeViA's effectiveness in enhancing segmentation performance by leveraging available text data, particularly in low-data regimes.

F. TeViA's Results under Different Label Percentages

We conducted experiments to evaluate TeViA's performance on the QaTa-COV19 dataset, using label percentages of 10%, 20%, 30%, 50%, 70% and 100%. As shown in Fig. 4, (1) TeViA surpasses the second-best method, LanGuideMedSeg, using only 50% of the annotations. This result validates TeViA's superiority in leveraging supplementary text data to enhance segmentation, showcasing more efficient knowledge extraction and integration. (2) With 10% paired image-text data, TeViA outperforms the top vision-only method, ACC-UNet, which uses 100% image data, by 1.57% in Dice score. This phenomenon underscores the significance of employing available complementary data to advance the primary segmentation task.

G. Visualization Analysis

Fig. 5 shows the features from the penultimate layer of the segmentation head, which were used to adjust the textual features during the experiments. It can be observed that integrating textual knowledge without proper alignment can result in significant false positive areas. However, by progressively applying the proposed T2V alignment $\mathcal{L}_{align}^{T2V}$ and the prototype attention Attn^{Proto}, the lesion regions become more distinctly highlighted.

In addition, Fig. 6 presents the segmentation results from all datasets. It is clear to see that our TeViA predicts lesion/cell regions more accurately than competing methods, particularly in the areas highlighted by the boxes. These visualizations qualitatively indicate the enhanced segmentation ability of TeViA by leveraging textual knowledge.

H. Discussion of Alignment Loss Function

We conducted a comprehensive evaluation of different alignment loss functions, including Cosine Similarity (default), KL Divergence, InfoNCE, and Wasserstein Distance, to assess

TABLE XII
EFFECT OF DIFFERENT SCALE WEIGHTS β_1, β_2, β_3 ON SEGMENTATION PERFORMANCE ON THE QATA-COV19 DATASET.

β_1	β_2	β_3	Dice (%)	mIoU (%)
0.13	0.33	0.53	90.08	81.95
0.23	0.33	0.43	90.55	82.73
0.33	0.33	0.33	**91.03**	**83.54**
0.43	0.33	0.23	90.62	82.85
0.53	0.33	0.13	90.24	82.22

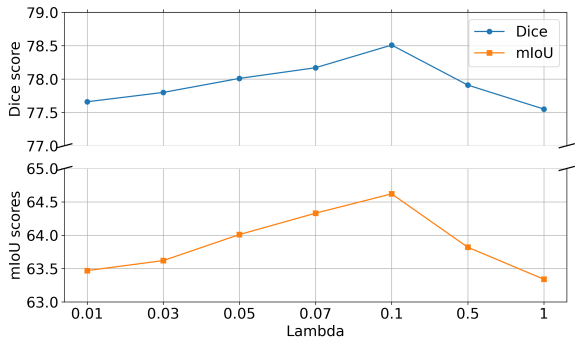


Fig. 7. Discussion of hyper-parameter λ on the MosMedData+ dataset.

their impact on the T2V alignment process. Our results on the QaTa-COV19 dataset show that Cosine Similarity achieves the best performance, attaining 91.03% Dice and 83.54% mIoU, outperforming other evaluated losses. While KL Divergence and Wasserstein Distance are effective in aligning cross-modal distributions, and InfoNCE promotes discriminative representations through contrastive learning, these alternatives often introduce increased computational complexity or exhibit heightened sensitivity to sampling strategies. In contrast, Cosine Similarity provides stable performance with lower computational overhead by directly enforcing angular closeness within the joint embedding space. These findings indicate that Cosine Similarity is a simple yet highly effective choice for cross-modal alignment within our framework.

I. Discussion of Hyperparameters

Table XII presents a hyperparameter analysis of the scale weights β_1, β_2, and β_3 on the QaTa-COV19 dataset. The results show that a balanced configuration (0.33, 0.33, 0.33) yields the best segmentation performance, achieving 91.03% Dice and 83.54% mIoU. Performance consistently declines as β_1 or β_3 deviates from this balanced setting, indicating the importance of maintaining equilibrium across scales. In particular, overemphasizing either coarse or fine features while keeping the middle scale fixed disrupts effective multi-scale context fusion. Nonetheless, the overall performance fluctuation across all settings remains within 1%, demonstrating that the model is not overly sensitive to these hyperparameters and exhibits stable performance. Similarly, the hyperparameter λ balances the primary segmentation loss and the T2V alignment loss. As shown in Figure 7, TeViA achieves optimal performance on the MosMedData+ dataset when λ is set to 0.1. A deviation from this value—either increasing or decreasing λ—leads to a modest decline in performance, suggesting that a relatively small weight for the alignment loss is favorable for segmentation. Nevertheless, the Dice score varies by only about 1% as λ changes from 0.01 to 1, further indicating that TeViA is robust to a wide range of λ values.

TABLE XIII
COMPARISONS WITH 3D MODELS ON THE MOSMEDDATA+.

Methods	MosMedData+	
	Dice (%)	mIoU (%)
nnUNet 3D	75.43	61.57
ACC-UNet 3D	74.68	59.59
TeViA	**78.49**	**64.59**

J. Comparisons with 3D Models

We compared TeViA with the 3D variants of nnUNet and ACC-UNet, two strong vision-only baselines. As shown in Table XIII, TeViA outperforms both methods on the MosMedData+ dataset, achieving 78.49% Dice and 64.59% mIoU, compared to 75.43% / 61.57% for nnUNet 3D and 74.68% / 59.59% for ACC-UNet 3D. These results demonstrate TeViA's effectiveness in leveraging textual information within a 2D framework, surpassing competitive 3D models and highlighting the complementary strength of text-guided supervision in medical image segmentation.

K. Discussion of Limitations and Application Scope

While our method achieves consistent gains across diverse datasets, several practical limitations must be acknowledged. First, its effectiveness relies on the availability of textual descriptions aligned with visual data. In our experiments, we used either curated annotations or automatically generated prompts. However, in clinical practice, such detailed annotations may be scarce, noisy, or inconsistent, potentially reducing the utility of text-guided supervision. Second, the current implementation is restricted to 2D inputs due to the slice-level structure of the available textual data. In most real-world 3D segmentation scenarios, only volume-level reports or coarse diagnostic summaries are available, lacking the granularity required for effective alignment. This mismatch may limit the method's direct applicability in volumetric contexts where fine-grained textual cues are not readily accessible.

VI. CONCLUSION

In this paper, we propose TeViA, a novel medical image segmentation method designed to fully leverage textual information through a segmentation-specific T2V alignment. This alignment strategy explicitly regularizes textual features, guiding them towards a distribution optimally suited for the segmentation task. Beyond achieving SOTA results, TeViA's broad applicability across diverse vision and text models is validated through comprehensive ablation studies. Furthermore, the demonstrated plug-and-play capability of our T2V alignment significantly boosts the performance of competing methods when integrated. Despite TeViA's promising results, the method may encounter specific limitations when dealing

with degraded image and text quality. Specifically, extreme noise levels or very low image contrast can notably impact the model's performance. These challenges primarily stem from the reduced reliability of visual cues, which can directly compromise the effectiveness of the T2V alignment process. Textual information at the decision-level, lesion-level, and location-level each contributes importantly to segmentation accuracy. Notably, location-level information exerts a particularly strong influence, while the simultaneous absence of multiple types of textual cues can lead to substantial performance degradation. To address these challenges related to both degraded image quality and noisy or incomplete textual information, future work will explore incorporating image denoising modules and contrast enhancement techniques as preprocessing steps to improve visual input quality prior to semantic alignment. Simultaneously, strategies aimed at enhancing textual robustness, such as text denoising and handling incomplete or corrupted cues, will be investigated. These strategies collectively aim to enhance the model's robustness for deployment in complex clinical environments.

REFERENCES

[1] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "A human-in-the-loop method for pulmonary nodule detection in ct scans," *Visual Intelligence*, vol. 2, no. 1, p. 19, 2024.

[2] Z. Lu, Y. Xie, Q. Zeng, M. Lu, Q. Wu, and Y. Xia, "Spot the difference: Difference visual question answering with residual alignment," in *MICCAI*, pp. 649–658, Springer, 2024.

[3] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, *et al.*, "Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset," *NeurIPS*, vol. 36, 2024.

[4] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, and P. Szczuko, "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Information Fusion*, vol. 90, pp. 316–352, 2023.

[5] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[6] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[7] S.-M. Park and Y.-G. Kim, "Visual language integration: A survey and open challenges," *Computer Science Review*, vol. 48, p. 100548, 2023.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, pp. 8748–8763, PMLR, 2021.

[9] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, p. ocae045, 2024.

[10] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue," in *AAAI*, vol. 38, pp. 19368–19376, 2024.

[11] Y. Zhong, M. Xu, K. Liang, K. Chen, and M. Wu, "Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images," in *MICCAI*, pp. 724–733, Springer, 2023.

[12] G.-E. Lee, S. H. Kim, J. Cho, S. T. Choi, and S.-I. Choi, "Text-guided cross-position attention for segmentation: Case of medical image," in *MICCAI*, pp. 537–546, Springer, 2023.

[13] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "Cxr-clip: Toward large scale chest x-ray language-image pre-training," in *MICCAI*, pp. 101–111, Springer, 2023.

[14] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *ICCV*, pp. 21372–21383, 2023.

[15] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *CVPR*, pp. 20730–20740, 2022.

[16] Y. Hu, Y. Chen, X. Xing, J. Zhang, B. M. Yerzhanuly, B. Matkerimqyzy, and Y. Xia, "Hyperbolic geometry-driven robustness enhancement for rare skin disease diagnosis," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[17] H. Wang, L. Huai, W. Li, L. Qi, X. Jiang, and Y. Shi, "Weakmedsam: Weakly-supervised medical image segmentation via sam with sub-class exploration and prompt affinity mining," *IEEE Transactions on Medical Imaging*, 2025.

[18] Q. Zeng, Y. Xie, Z. Lu, M. Lu, Y. Wu, and Y. Xia, "Segment together: A versatile paradigm for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, 2025.

[19] Y. Xie, J. Zhang, Y. Xia, and Q. Wu, "Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier," in *ECCV*, pp. 558–575, Springer, 2022.

[20] Y. Chen, X. Guo, Y. Xia, and Y. Yuan, "Disentangle then calibrate with gradient guidance: A unified framework for common and rare disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1816–1827, 2024.

[21] Z. Qian, K. Li, M. Lai, E. I.-C. Chang, B. Wei, Y. Fan, and Y. Xu, "Transformer based multiple instance learning for weakly supervised histopathology image segmentation," in *MICCAI*, pp. 160–170, Springer, 2022.

[22] Q. Zeng, Z. Lu, Y. Xie, and Y. Xia, "Pick: Predict and mask for semi-supervised medical image segmentation," *International Journal of Computer Vision*, pp. 1–16, 2025.

[23] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, *et al.*, "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, 2022.

[24] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[25] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *ECCV*, pp. 206–222, Springer, 2016.

[26] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, 2017.

[27] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[28] Z. Jia, X. Huang, I. Eric, C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE Transactions on Medical Mmaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[29] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang, *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, p. 101832, 2021.

[30] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI*, pp. 556–564, Springer, 2015.

[31] S. P. Morozov, A. E. Andreychenko, N. Pavlov, A. Vladzymyrskyy, N. Ledikhova, V. Gombolevskiy, I. A. Blokhin, P. Gelezhe, A. Gonchar, and V. Y. Chernina, "Mosmeddata: Chest ct scans with covid-19 related findings dataset," *arXiv preprint arXiv:2005.06465*, 2020.

[32] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[33] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[34] X. Ma, H. Cui, S. Li, Y. Yang, and Y. Xia, "Deformable medical image registration with global–local transformation network and region similarity constraint," *Computerized Medical Imaging and Graphics*, vol. 108, p. 102263, 2023.

[35] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 143–154, 2022.

[36] Q. Zeng, J. Geng, W. Jiang, K. Huang, and Z. Wang, "Idln: Iterative distribution learning network for few-shot remote sensing image scene

classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
[37] Q. Zeng, J. Geng, K. Huang, W. Jiang, and J. Guo, "Prototype calibration with feature generation for few-shot remote sensing image scene classification," *Remote Sensing*, vol. 13, no. 14, p. 2728, 2021.
[38] H. Hassan, Z. Ren, C. Zhou, M. A. Khan, Y. Pan, J. Zhao, and B. Huang, "Supervised and weakly supervised deep learning models for covid-19 ct diagnosis: A systematic review," *Computer Methods and Programs in Biomedicine*, p. 106731, 2022.
[39] X. Ma, S. Xu, J. Zhou, Q. Yang, Y. Yang, K. Yang, and S. H. Ong, "Point set registration with mixture framework and variational inference," *Pattern Recognition*, vol. 104, p. 107345, 2020.
[40] S. Zhang, J. Zhang, Y. Xie, and Y. Xia, "Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation," in *MICCAI*, pp. 109–118, Springer, 2023.
[41] Q. Zeng, Z. Lu, Y. Xie, M. Lu, X. Ma, and Y. Xia, "Reciprocal collaboration for semi-supervised medical image classification," in *MICCAI*, pp. 522–532, Springer, 2024.
[42] Y. Chen, M. Mancini, X. Zhu, and Z. Akata, "Semi-supervised and unsupervised deep visual learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[43] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training," in *CVPR*, pp. 15671–15680, 2023.
[44] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature Medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.
[45] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 863–874, 2024.
[46] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *EMNLP*, pp. 3876–3887, 2022.
[47] Z. Lu, Q. Zeng, M. Lu, G. Chen, and Y. Xia, "Bridging the semantic gap in medical visual question answering with prompt learning," *IEEE Transactions on Medical Imaging*, 2025.
[48] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "Radbert: adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, vol. 4, no. 4, p. e210258, 2022.
[49] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
[50] M. Yasunaga, J. Leskovec, and P. Liang, "Linkbert: Pretraining language models with document links," in *ACL*, pp. 8003–8016, 2022.
[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[52] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.
[53] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
[54] I. Hartsock and G. Rasool, "Vision-language models for medical report generation and visual question answering: A review," *arXiv preprint arXiv:2403.02469*, 2024.
[55] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.
[56] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "Clip-driven universal model for organ segmentation and tumor detection," in *ICCV*, pp. 21152–21164, 2023.
[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[58] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *ECCV*, pp. 1–21, Springer, 2022.
[59] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.
[60] A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. Gabbouj, "Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images," in *ICIP*, pp. 2306–2310, IEEE, 2022.
[61] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.
[62] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre, "Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network," *Biomedical Signal Processing and Control*, vol. 61, p. 102027, 2020.
[63] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "Medclip-sam: Bridging text and image towards universal medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 643–653, Springer, 2024.
[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *NeurIPS*, vol. 32, 2019.
[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, Springer, 2015.
[66] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *MICCAIW*, pp. 3–11, Springer, 2018.
[67] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.
[68] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[69] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *ECCVW*, pp. 205–218, Springer, 2022.
[70] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *AAAI*, vol. 36, pp. 2441–2449, 2022.
[71] N. Ibtehaz and D. Kihara, "Acc-unet: A completely convolutional unet model for the 2020s," in *MICCAI*, pp. 692–702, Springer, 2023.
[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.
[73] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, pp. 10012–10022, 2021.
[74] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, pp. 11976–11986, 2022.
[75] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*, pp. 2–25, PMLR, 2022.
[76] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *ICCV*, pp. 3942–3951, 2021.
[77] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *CVPR*, pp. 18155–18165, 2022.
[78] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[79] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE*, pp. 248–255, Ieee, 2009.
[80] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, pp. 2613–2622, 2021.
[81] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *CVPR*, pp. 11514–11524, IEEE, 2023.
[82] Q. Zeng, Y. Xie, Z. Lu, M. Lu, J. Zhang, Y. Zhou, and Y. Xia, "Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024.