TP-UNet: Temporal Prompt Guided UNet for Medical Image Segmentation

Anonymous ACL submission

Abstract

The advancement of medical image segmentation techniques has been propelled by the adoption of deep learning techniques, particularly UNet-based approaches, which exploit semantic information to improve the accuracy of segmentations. However, the order of organs in scanned images has been disregarded by current medical image segmentation approaches based on UNet. Furthermore, the inherent network structure of UNet does not provide direct capabilities for integrating temporal information. To efficiently integrate temporal information, we propose TP-UNet that utilises temporal prompts, encompassing organ-construction relationships, to guide the segmentation UNet model. Specifically, our framework is featured with cross-attention and semantic alignment based on unsupervised contrastive learning to combine temporal prompts and image features effectively. Extensive evaluations on two medical image segmentation datasets demonstrate the state-of-the-art performance of TP-UNet. The anonymous codebase is available at `https://anonymous.4open.science/r/TP_UNet`.

1 Introduction

Medical image segmentation holds a pivotal position within the realm of modern medicine, playing a fundamental role in disease diagnosis, surgical planning, and treatment monitoring (Khened et al., 2018). The primary objective of this task is to accurately separate and label distinct structures or tissues depicted in medical images, enabling healthcare professionals to conduct meticulous analysis and achieve precise diagnoses. Notably, with the advancements in deep learning techniques, certain networks built upon UNet and its variants have exhibited commendable segmentation accuracy by leveraging semantic information extracted from medical images (Siddique et al., 2021). Although promising results were reported, existing UNet-based approaches lack consideration for the temporal information present in scanned medical images (Ye et al., 2022). To better understand the temporal information, we have visualised it in Fig. 1. Incorporating temporal information, which represents a sequence of medical images, has the potential to enhance the accuracy of medical segmentation. For instance, in a series of N slices from a patient, denoted as N_i^{th}/N. Notably, certain organs such as the stomach, large intestine, and small intestine exhibit specific temporal patterns within a given interval $(N_1^{th}/N, N_m^{th}/N)$, often following a normal distribution. Specifically, their normal distributions can be expressed as $\mathcal{N}_{stomach}(\mu_{stomach}, \sigma_{stomach})$, $\mathcal{N}_{large}(\mu_{large}, \sigma_{large})$ and $\mathcal{N}_{small}(\mu_{small}, \sigma_{small})$, respectively. In imaging modalities like MRI or CT scans, where the imaging process typically progresses from top to bottom, the stomach predominantly appears in the early to mid-time intervals, while the small intestine is more prevalent in the mid-time interval, and the large intestine is primarily observed during the late-time interval. Consequently, we have $\mu_{stomach} \leq \mu_{small} \leq \mu_{large}$. Thus, within a specific temporal interval, it is essential to prioritise stomach segmentation during the early stages and focus on large intestine segmentation in the later stages to improve the performance of medical segmentation models. Despite the evident importance of temporal information in enhancing segmentation accuracy, its consideration is often overlooked in current research.

To exploit the temporal information inherent in medical images, we propose TP-UNet, a framework that leverages temporal prompts to guide the learning process of the UNet model. The temporal prompt offers textual signals for guiding the segmentation model in learning semantic and sequential information from medical images. These textual signals first undergo high-dimensional embedding via a well-trained encoder and then interact with the feature map from the image encoder. Due to the different encoding processes of text and im-

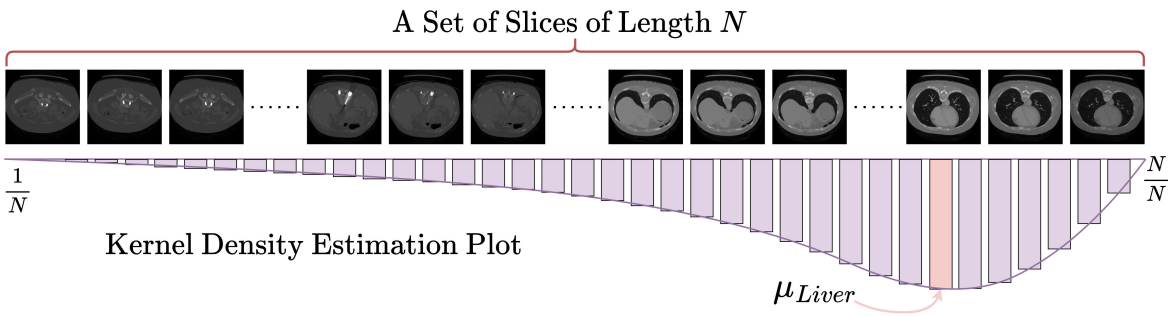

Figure 1: **The temporal information of liver.** We visualized the temporal information of the liver. From the kernel density plot of the liver occurrence probability, it can be seen that the distribution approximately follows a normal distribution $\mathcal{N}(\mu_{Liver}, \sigma_{Liver})$ for a set of timestamps ranging from $\frac{1}{N}$ to $\frac{N}{N}$. The timestamp with the highest frequency of liver occurrence is approximately 0.78. For multiple organs, such as the three organs in the UW-Madison dataset (i.e., stomach, large intestine, and small intestine), their probabilities of occurrence at different timestamps also vary, typically $\mu_{stomach} \leq \mu_{small} \leq \mu_{large}$. This temporal information is crucial for guiding the model in segmentation tasks.

age, simply using a linear mapping between image embeddings and textual embeddings for interaction may lead to suboptimal fusion results or even degrade model performance (Li et al., 2021). To address this, we perform a semantic alignment operation before the interaction between text and image modalities, utilizing unsupervised contrastive learning for text representation and image representation to narrow the semantic gap. Finally, for modality fusion, we employ a cross-attention mechanism to aggregate the aforementioned updated text representation and image representation. This process yields a unified representation that serves as input to the decoder of the UNet model. Our main contributions can be summarized as follows:

- We propose TP-UNet, a simple and effective framework for medical image segmentation, which can guide the segmentation model to learn the temporal information in medical images through textual prompts.

- We propose a two-stage process of semantic alignment and modal fusion to narrow the semantic gap between temporal prompts and image features and effectively aggregate them into a unified representation.

- We conducted extensive experiments on two medical image segmentation datasets, including the LITS 2017 dataset and the UW-Madison dataset. The results of the experiments demonstrated that our method achieved a new state-of-the-art (SOTA) performance.

2 Related Work

2.1 Prompt Learning

Prompt learning is a crucial research area in the field of natural language processing (NLP). It focuses on designing effective prompts or questions to guide models in generating accurate and relevant outputs for specific tasks. By tailoring prompts to the task at hand, prompt learning helps the model concentrate on essential information and reduces the search space, thus improving model performance. This technique has proven successful in various NLP tasks, including text classification, named entity recognition, and machine translation. Recently, this approach has been applied to medical segmentation, aiming to enhance the segmentation capabilities of models through text prompts. Jie Liu et al. constructed prompts based on the names of organs that need to be segmented (Liu et al., 2023). Junde Wu et al., on the other hand, built prompts based on textual descriptions of organs, including their functions, shapes, and appearances, achieving significant improvements in segmentation performance (Wu et al., 2023). However, it is evident that neither of these approaches utilized temporal information from medical images. In this study, we designed prompts based on the temporal information of medical images, aiming to guide the segmentation model using temporal information for better performance.

2.2 Multimodal Contrastive Learning

Multimodal contrastive learning is a powerful technique in the field of multimodal learning, which considers multiple modalities such as text, images,

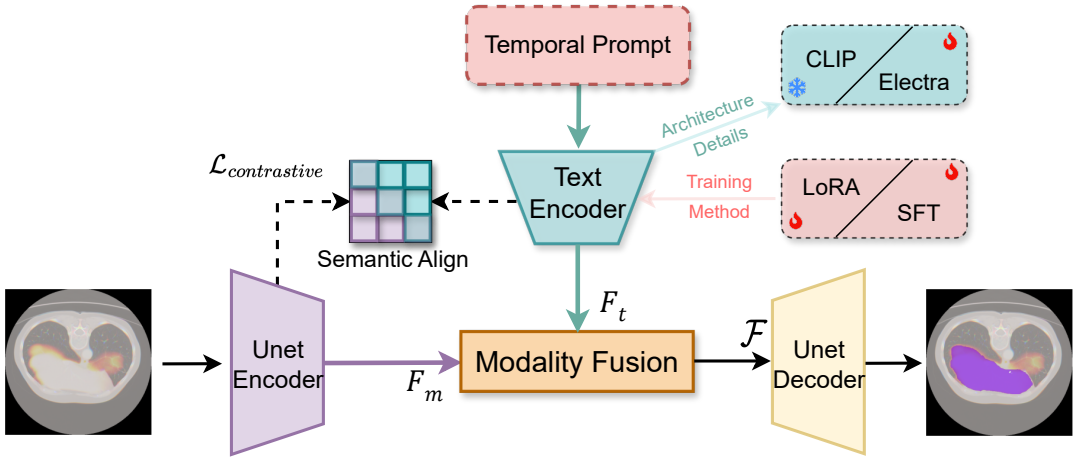

Figure 2: **The general framework of TP-UNet.** For a given medical image I that needs segmentation, TP-UNet first automatically generates its corresponding temporal prompt P_t. The UNet encoder then extracts features from the input medical image I. These extracted features are fused with the encoded temporal prompt F_t. Prior to fusion, a semantic alignment operation is performed to bridge the gap between different modality encoders. Finally, the UNet decodes the fused features to produce the final masks. It should be noted that the text encoder in this study employs two architectures: CLIP and Electra. These architectures are trained using the LoRA and SFT methods, respectively.

and audio. The goal of this method is to learn meaningful representations by maximizing the similarity between samples from the same category and minimizing the similarity between samples from different categories. In the context of multimodal learning, this involves aligning the representations of different modalities in a shared embedding space to capture their relationships and interactions. Multimodal contrastive learning leverages complementary information from multiple modalities, enhancing the model's understanding of the overall data and its representation learning capabilities. This approach has been widely applied in the medical domain. Yuhao Zhang et al. proposed ConVIRT (Zhang et al., 2022), which uses a bidirectional contrastive objective function to maximize the consistency between true matches and random pairs of images and texts, achieving unsupervised training. This method leverages paired text data across domains without requiring additional expert input. In image classification tasks, ConVIRT achieves high data efficiency, as it can achieve comparable or even better performance than models initialized with ImageNet using only 10% labeled training data. Shih-Cheng Huang et al. proposed the GLoRIA (Huang et al., 2021) framework, which maximizes the correlation between medical images and texts using global and local contrastive loss functions, leading to improved performance in downstream tasks. In summary, multimodal contrastive learning can enhance data efficiency, align semantic information between different modalities, and improve performance in downstream tasks. The contrastive learning used in this paper addresses the issue of similarity between different modalities, resulting in better generalization and segmentation performance for downstream medical image segmentation tasks.

3 Methodology

In this section, we introduce the TP-UNet model (as shown in Fig 2), which addresses the issue of temporal information forgetting in medical image analysis by designing the Temporal Prompt module. Additionally, we utilize the Semantic Align module to bridge the semantic gap between temporal prompt and image modalities. The combined effect of these two key components significantly enhances the performance of TP-UNet in medical image segmentation, enabling more precise and consistent segmentation of dynamic images.

3.1 Temporal Prompt

Temporal information plays a crucial role in improving model segmentation performance. we devised a set of prompts to guide the model in understanding the temporal information of medical images. In this study, temporal information is repre-

sented as N_i^{th}/N, indicating that this information is mapped to the interval [0, 1]. The occurrence probability of organs follows a normal distribution within this interval, allowing the model to comprehend the varying probabilities of organ appearance at different timestamps, thereby adjusting its focus on different organs accordingly. The temporal prompt template defined in this study is as follows: *"This is {an MRI / a CT} of the {organ} with a segmentation period of {N_i^{th}/N}."* Here, the type of medical image and the organ can be selected, while N is determined by the size of a set of slices. In this study, the temporal prompts are automatically generated. Before being input into TP-UNet, a set of prompts is automatically created using the numpy and pandas libraries based on the type of image selected by the physician. In specific situations, radiologists can also choose the range of timestamps for segmentation by dragging to select the desired range. This allows TP-UNet to save a significant amount of time during inference by generating prompts only for the specified range of slices. The generation of a temporal prompt for a single set of slices takes less than 1ms, which is highly significant for clinical applications in radiology.

3.2 Multimodal Encoder

We first define the input medical image as I and the generated temporal prompt as P_t. For the text-based temporal prompt and the medical image requiring segmentation, we designed a multimodal encoder. For the input text modality P_t, we adopted two encoding methods. The first method utilizes the popular multimodal text encoder CLIP (Radford et al., 2021). While CLIP performs well with general natural language, directly applying it to medical text may introduce a domain gap. Therefore, we employed parameter-efficient fine-tuning (PEFT) (Ding et al.) using the LoRA (Hu et al., 2021) method to adapt CLIP more effectively to our task. The second text encoder we used is Electra (Clark et al., 2020), another popular text encoder. We compared the performance of these two encoders in the experiments section. We performed supervised fine-tuning (SFT) (Ouyang et al., 2022) on the pre-trained Electra model. Both fine-tuned text encoders demonstrated good performance.

For the medical image modality, we used the conventional UNet method for segmentation. We integrated the low-level semantics extracted by UNet with the temporal prompt to guide the model for more effective segmentation based on temporal information. The details of the fusion method will be presented in Section 3.4.

3.3 Semantic Align

In this context, we define the encoded image feature $F_m \in \mathbb{R}^{B\times C\times H\times W}$ and the textual temporal prompt encoded feature $F_t \in \mathbb{R}^{B\times L\times D}$. Before modality fusion, I undergoes an UNet encoder block, while P_t via the text encoder. Due to the disparate network architectures of the two models, they originate from different semantic spaces, potentially leading to a performance decrease after fusion. Therefore, it becomes essential to align the semantics of F_m and F_t before modality fusion. To achieve this, we introduce a semantic align module, aiming to bring semantically similar pairs of F_{mi} and F_{ti} closer together in a batch, while pushing semantically dissimilar F_{mi} and non-corresponding F_{tj} further apart. Consequently, the first contrastive loss function is an image-to-text contrastive loss for the i-th pair:

$$\ell_i^{(F_m\to F_t)} = -\log \frac{\exp(\langle F_{mi}, F_{ti}\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle F_{mi}, F_{tk}\rangle/\tau)} \tag{1}$$

$$\langle F_{mi}, F_{ti}\rangle = F_{mi}^{\top} F_{ti}/\|F_{mi}\|\|F_{ti}\| \tag{2}$$

where $\tau \in \mathbb{R}^+$ represents a temperature parameter. The second loss function is a text-to-image contrastive loss for the i-th pair:

$$\ell_i^{(F_t\to F_m)} = -\log \frac{\exp(\langle F_{ti}, F_{mk}\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle F_{ti}, F_{mk}\rangle/\tau)} \tag{3}$$

In the end, the loss we need to optimize is:

$$\mathcal{L}_{contrastive} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda\ell_i^{(F_m\to F_t)} + (1-\lambda)\ell_i^{(F_t\to F_m)}\right) \tag{4}$$

where $\lambda \in [0, 1]$ is a scalar weight. Through the semantic alignment module, the semantic representations of different modalities are aligned. This lays a solid foundation for the subsequent modality fusion.

3.4 Modality Fusion

Temporal prompt is crucial to improve the segmentation performance of the model. Consequently, there should be increased emphasis on the design of modality fusion between the temporal prompt modality and the visual modality. Therefore, we designed the cross-attention mechanism, which can be represented as follows:

$$\mathcal{F} = \text{softmax}\left(\frac{\left([F'_m;F'_t]\mathbf{W}^Q\left([F'_m;F'_t]\mathbf{W}^K\right)^\top\right)}{\sqrt{d_k}}\right)[F'_m;F'_t]\mathbf{W}^V \tag{5}$$

Where $[;]$ represents the concatenate operation, $\mathcal{F}$ is the pixel-wise attention map, and d_k is a scaling factor. F'_m and F'_t are the projections of F_m and F_t. $\mathbf{W}^Q$, $\mathbf{W}^K$ and $\mathbf{W}^V$ are the corresponding weight matrices. Finally, the feature map $\mathcal{F}$ is concatenated with the first-level skip-connection feature map of the UNet. It undergoes a convolutional layer and a ReLU activation function before passing through a 1×1 convolutional layer to generate the final segmentation image.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

UW-Madison Dataset Collection: [1] The dataset originates from MRI scan images of multiple patients at the Carbone Cancer Center of the University of Wisconsin-Madison. It primarily consists of MRI images of the colon and stomach regions. We call it the UW-Madison dataset (happyharrycn et al., 2022) and divide it according to a 7: 1: 2 ratio into training, validation, and testing set. Among them, the training set contains 26746 images, the validation set contains 3820 images, and the testing set contains 7642 images.

LITS Dataset Collection: [2] LITS (Bilic et al., 2023) is an acronym for Liver Tumor Segmentation Benchmark. The data and segmentations are provided by various clinical sites around the world. The dataset contains CT scans of 130 patients. But these scans are 3D nii files, what we need are 2D slices. We divide the LITS dataset into 58638 2D slices, but a large number of 2D slices are also redundant. We finally selected 10967 2D slices with sequence information, and divided these slices according to a 7: 1: 2 ratio into training, validation, and testing set.

4.1.2 Implementation details

We choose the Adaptive Momentum Estimation with a weight decay of 0.000001 as the training optimizer. Meanwhile, the initial learning rate is 0.00003, and the weights change with the cosine annealing learning rate; the initial temperature is 25, and the maximum temperature is 96.875. We use the PyTorch training framework and some data augmentation methods, such as CoarseDropout, HorizontalFlip, and ShiftScaleRotate. The loss function uniformly adopts the average value of Binary Cross-Entropy ($\mathcal{L}_{BCE}$) and Tversky loss ($\mathcal{L}_{Tversky}$), $\mathcal{L}_{Tversky}$ is the loss function to solve imbalanced classification problems.

4.1.3 Evaluation metrics

We use the Jaccard coefficient and the Dice coefficient to evaluate the performance of the model, which can measure the performance of the model by calculating the similarity between the ground-truth annotation and the predicted annotation. Their calculation can be expressed as follows:

$$\begin{aligned} Jaccard &= \frac{A\cap B}{A\cup B}, \\ Dice &= \frac{2(A\cap B)}{A+B}, \end{aligned} \tag{6}$$

where A and B are binary matrices representing the ground-truth annotation and the predicted annotation, respectively.

4.2 Comparison with the Baselines

To demonstrate the effectiveness of TP-UNet, we conducted extensive experiments on two different datasets. Several commonly used medical image segmentation models are selected for experimental comparison. We use the Jaccard coefficient and the Dice coefficient to evaluate the performance.

As shown in Table 1, TP-UNet achieved the best performance in all three organ categories (large intestine, small intestine, and stomach) as well as the overall average performance on the UW-Madison dataset. Compared to UNet, the Dice score improved by an average of 4.44%, with the most significant improvement of 5.32% in the Small Intestine category. In addition to UNet, we also compared our TP-UNet with several other methods listed in Table 1. TP-UNet outperformed the current state-of-the-art (Swin UNet (Cao et al., 2023)) by 1.3% in the Dice score, with the most significant improvement of 1.7% in the Small Intestine category.

We also conducted experiments on the LITS 2017 dataset. As shown in Table 2, our method achieved the highest Dice and Jaccard scores for liver segmentation. Compared to UNet, the Dice score for the liver increased by 6.08%, and the Jaccard score increased by 6.33%. On the LITS 2017 dataset, TP-UNet outperformed the current state-of-the-art method by 9.21% in Dice score, with

[1] https://www.kaggle.com/datasets/uwmgi-mask-dataset
[2] https://competitions.codalab.org/competitions/LITS 2017

Model	Backbone	Large Intestine		Small Intestine		Stomach		Average	
		Dice ↑	Jacc ↑	Dice ↑	Jacc ↑	Dice ↑	Jacc ↑	Dice ↑	Jacc ↑
UNet (Ronneberger et al., 2015)	VGG16 (Simonyan and Zisserman, 2014)	0.8709	0.8329	0.8546	0.8187	0.9210	0.9001	0.8822	0.8506
UNet (Ronneberger et al., 2015)	Resnet50 (He et al., 2016)	0.8621	0.8211	0.8487	0.8122	0.9106	0.8880	0.8738	0.8405
UNet++ (Zhou et al., 2019)	VGG16 (Simonyan and Zisserman, 2014)	0.8825	0.8463	0.8738	0.8394	0.9394	0.9201	0.8986	0.8686
UNet++ (Zhou et al., 2019)	Resnet50 (He et al., 2016)	0.8613	0.8223	0.8537	0.8172	0.9178	0.8970	0.8776	0.8455
Attention UNet (Oktay et al., 2018)	VGG16 (Simonyan and Zisserman, 2014)	0.8925	0.8580	0.8857	0.8519	0.9472	0.9292	0.9085	0.8797
scSE UNet (Roy et al., 2018)	VGG16 (Simonyan and Zisserman, 2014)	0.8909	0.8561	0.8750	0.8398	0.9491	0.9312	0.9050	0.8757
Trans UNet (Chen et al., 2021)	Vision Transformer (Dosovitskiy et al., 2020)	0.8983	0.8597	0.8847	0.8543	0.9356	0.9185	0.9062	0.8775
Swin UNet (Cao et al., 2023)	Swin Transformer (Liu et al., 2021)	0.9034	0.8704	0.8908	0.8617	0.9467	0.9279	0.9136	0.8819
TP-UNet $_{Electra}$	VGG16 (Simonyan and Zisserman, 2014)	0.9170	0.8781	0.9078	0.8704	0.9551	0.9343	0.9266	0.8943
TP-UNet $_{CLIP}$	VGG16 (Simonyan and Zisserman, 2014)	**0.9190**	**0.8810**	**0.9102**	**0.8727**	**0.9566**	**0.9358**	**0.9286**	**0.8965**

Table 1: Main results on the UW-Madison dataset. We compare our proposed TP-UNet (with Electra and CLIP text encoders) against several baseline segmentation models, including U-Net, U-Net++, Attention U-Net, scSE U-Net, TransUNet, and Swin UNet, under different backbones (VGG16, ResNet50, ViT, Swin Transformer). Performance is measured by Dice score and Jaccard index for each organ category (Large Intestine, Small Intestine, Stomach), as well as the average across three organs. All metrics are presented as decimals, with ↑ indicating that higher values are better. Our TP-UNet with CLIP backbone consistently achieves the highest scores across all evaluation metrics, with the best results highlighted in bold.

Model	BackBone	Liver	
		Dice ↑	Jacc ↑
UNet	VGG16	0.8517	0.8147
UNet	Resnet50	0.7223	0.6848
UNet++	VGG16	0.8486	0.8098
UNet++	Resnet50	0.8135	0.7751
Attention UNet	VGG16	0.8500	0.8137
scSE UNet	VGG16	0.8498	0.8143
Trans UNet	Vision Transformer	0.8069	0.7699
Swin UNet	Swin Transforme	0.8204	0.7833
TP-UNet $_{Electra}$	VGG16	**0.9125**	**0.8780**
TP-UNet $_{CLIP}$	VGG16	0.8657	0.8269

Table 2: Main results on the LITS dataset. Compared with several U-Net variants and Transformer-based models, our TP-UNet$_{\text{Electra}}$ attains the best segmentation performance on the liver.

the most significant improvement of 9.47% in the Small Intestine category.

4.3 Case Study

To further evaluate the effectiveness of TP-UNet, we conducted a qualitative comparison on representative samples from the LiTS2017 dataset, as shown in Figure 3. The figure presents the original CT images, ground-truth annotations, and segmentation results generated by TP-UNet, U-Net, and TransUNet. TP-UNet consistently produces segmentation masks that are visually closer to the ground truth.

Compared with U-Net, TP-UNet better captures complex liver boundaries and preserves fine structural details. U-Net often generates over-smoothed contours and inaccurate predictions in ambiguous or low-contrast regions. Although TransUNet benefits from transformer-based global feature modeling, it occasionally suffers from incomplete segmentation and inaccurate edge localization. In contrast, TP-UNet effectively integrates local spatial information with global contextual dependencies, enabling more robust feature representation.

In challenging cases with irregular liver shapes and blurred boundaries, TP-UNet produces more complete and accurate segmentation results while reducing false predictions. These qualitative results further demonstrate the robustness and effectiveness of the proposed architecture.

4.4 Ablation Study

We conducted ablation experiments on the UW-Madison dataset and the LITS2017 dataset to demonstrate the effectiveness of the TP-UNet methods.

The experimental results are shown in Table 3, where a lower score indicates a greater contribution of the module to the TP-UNet model. First, we verified the effectiveness of the temporal information by removing the timestamp from the temporal prompt, fixing the prompt template to *"This is an MRI / a CT of the organ"*. Keeping other settings unchanged, we observed a 2.1% decrease in the mDice score on the UW-Madison dataset. This

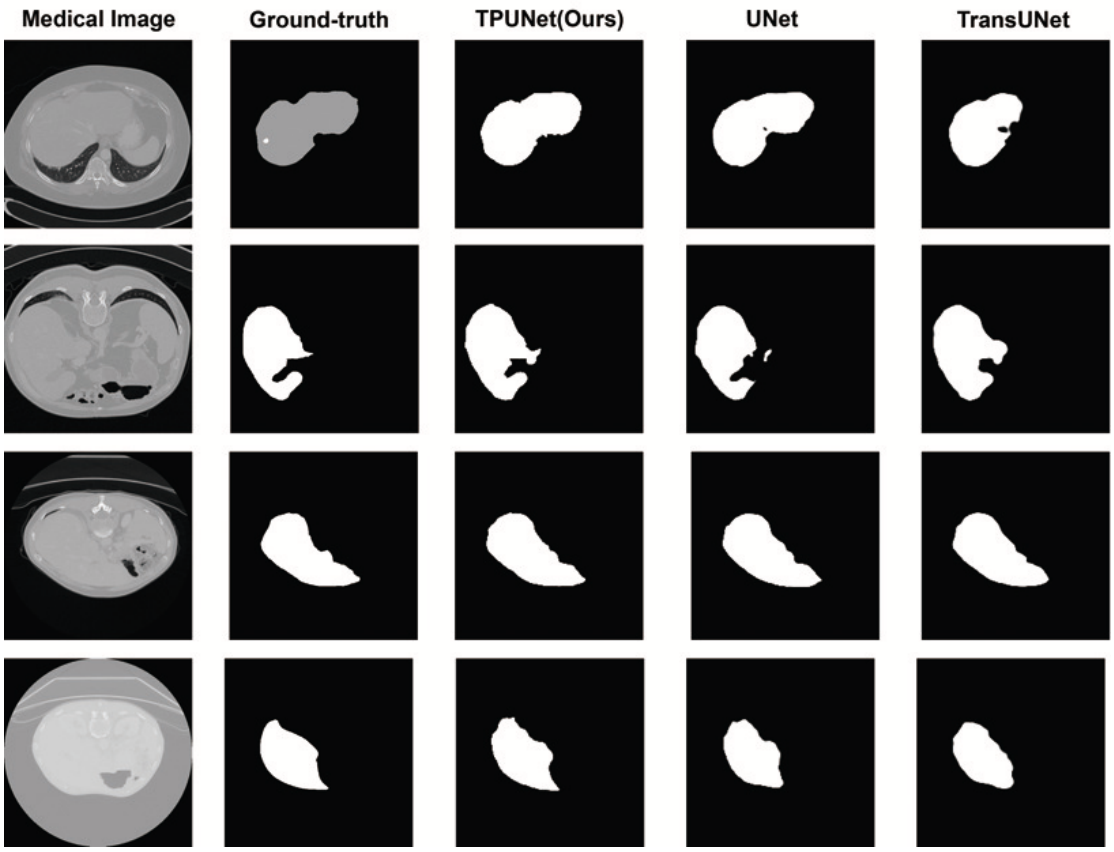

Figure 3: **Case Study.** Qualitative comparison on four samples from the LITS dataset. For each case, we show the input medical image, ground-truth mask, and segmentation results of our TP-UNet, U-Net, and TransUNet. Our method produces masks that are visually closer to the ground-truth, demonstrating superior segmentation quality.

further demonstrates the effectiveness of incorporating temporal information, which significantly helps guide the model in enhancing segmentation performance.

Next, we removed the entire temporal prompt and did not use a text encoder. Consequently, the modality fusion changed to a self-attention mechanism. From the results, we observed a significant decrease of 5.36% in the mDice score on the LITS dataset. This demonstrates that the temporal prompt not only provides valuable temporal information but also that the selected organ and image type are beneficial for segmentation.

We also explored the effectiveness of the semantic alignment module. We performed modality fusion directly without semantic alignment beforehand. The results showed that the mDice score on the UW-Madison dataset decreased by 1.01%. This demonstrates that semantic alignment is essential for multimodal fusion, as it helps reduce the domain gap between different modality encoders. This improvement enhances the efficiency of multimodal fusion and the overall performance of the model.

We also investigated the performance of the modality fusion module. We replaced the modality fusion with a direct concatenation of the inputs before the final segmentation decoder. The results indicated that the modality fusion module is indispensable for the TP-UNet framework. The proposed modality fusion method in this paper demonstrates both efficiency and superior performance.

Through the above four sets of experiments, we can identify the effectiveness of the Temporal Prompt, the necessity of the semantic alignment module, and the efficiency of the modality fusion module. These three components are indispensable for TP-UNet and are crucial for its superior performance.

4.5 Generalization Experiments

To further evaluate the robustness and transferability of the proposed TP-UNet, we conducted generalization experiments on the LiTS dataset using different backbone settings and compared the results with several recent state-of-the-art methods,

Model	UW-Madison Dataset								LITS 2017	
	Large Intestine		Small Intestine		Stomach		Average		Liver	
	Dice ↓	Jacc ↓	Dice ↓	Jacc ↓	Dice ↓	Jacc ↓	Dice ↓	Jacc ↓	Dice ↓	Jacc ↓
TP-UNet $_{Electra}$	0.9170	0.8781	0.9078	0.8704	0.9551	0.9343	0.9266	0.8943	0.9125	0.8780
w/o Temporal Information	0.8924	0.8545	0.8787	0.8423	0.9306	0.9117	0.9048	0.8733	0.8735	0.8396
w/o Temporal Prompt	0.8797	0.8427	0.8666	0.8305	0.9306	0.9117	0.8923	0.8617	0.8589	0.8246
w/o Semantic Align	0.9082	0.8683	0.8958	0.8586	0.9455	0.9227	0.9165	0.8832	0.9053	0.8681
w/o Modality Fusion	0.9101	0.8684	0.8996	0.8623	0.9434	0.9237	0.9177	0.8848	0.9047	0.8709

Table 3: Ablation study on the UW-Madison and LITS datasets. We evaluate the contribution of key components in TP-UNet: temporal information, temporal prompt, semantic alignment, and modality fusion. Removing any component leads to performance drops, demonstrating the effectiveness of each module.

Model	Reference	Liver	
		Dice ↑	Jacc ↑
CSCPA	CVPR2025	0.8286	0.7829
SKCDF	CVPR2025	0.8301	0.7849
CSCPA+TP	CVPR2025	0.8344	0.7882
SKCDF+TP	CVPR2025	0.8536	0.8116
TP-UNet $_{CLIP}$	–	0.8657	0.8269
TP-UNet $_{Electra}$	–	**0.9125**	**0.8780**

Table 4: Generalization experiments on the LITS dataset. Compared with recent methods (CSCPA, SKCDF) published at CVPR 2025, our TP-UNet with Electra backbone achieves significantly higher Dice and Jaccard scores, demonstrating strong generalization.

as shown in Table 4. Specifically, we compared our approach against CSCPA (Ding et al., 2025) and SKCDF (Zhang et al., 2025), two representative methods published at CVPR 2025.

The experimental results demonstrate that integrating the proposed TP module consistently improves segmentation performance across different baseline architectures. For example, after incorporating TP into CSCPA and SKCDF, the Dice scores increase from 0.8286 to 0.8344 and from 0.8301 to 0.8536, respectively. Similar improvements can also be observed in the Jaccard metric, indicating that the proposed module enhances the feature representation capability and improves segmentation consistency.

Furthermore, TP-UNet equipped with different text encoders achieves superior performance compared with all competing methods. TP-UNet$_{CLIP}$ obtains a Dice score of 0.8657 and a Jaccard score of 0.8269, already outperforming recent transformer-based approaches. More importantly, TP-UNet$_{Electra}$ achieves the best overall performance, reaching a Dice score of 0.9125 and a Jaccard score of 0.8780. These results significantly surpass those of existing methods, demonstrating the effectiveness of combining semantic prior information with the TP-UNet architecture.

Overall, the experimental results confirm that TP-UNet possesses strong generalization ability across different backbone networks and challenging liver segmentation scenarios. The consistent performance gains suggest that the proposed framework can effectively capture both local structural information and global semantic context, leading to more accurate and robust medical image segmentation.

5 Conclusion

In this paper, we propose a temporal prompt guiding framework for medical image segmentation, which guides the learning of a segmentation model to the inherent temporal information of scanned images through straightforward temporal prompts. In addition, we further propose a two-stage process including semantic alignment and modality fusion to aggregate temporal prompts textual representation and image representations via multimodal contrast learning and cross-attention mechanisms. Our proposed framework also validates the necessity of temporal information in medical image segmentation tasks through promising results, consistently outperforming existing methods and demonstrating strong generalization across different datasets and organ types. In future work, we will extend the proposed framework to more complex scenarios, such as multi-organ segmentation with partial annotations and weakly supervised settings.

Limitations

TP-UNet is designed for 2D slice-wise segmentation and does not explicitly capture 3D volumetric information. Consequently, it may produce inconsistent predictions across adjacent slices in CT or MRI volumes, and it cannot leverage inter-slice contextual cues. Additionally, the temporal prompt relies on a normalized slice index, which presumes a consistent scanning order and organ distribution—conditions that may not always hold in clinical practice. Future work will extend the framework to handle 3D inputs and more flexible temporal priors.

References

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, and 1 others. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2023. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer.

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

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Mi Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hatao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models.

Zhenhui Ding, Guilian Chen, Qin Zhang, Huisi Wu, and Jing Qin. 2025. Csc-pa: Cross-image semantic correlation via prototype attentions for single-network semi-supervised breast tumor segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15632–15641.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

happyharrycn, Maggie, Phil Culliton, Poonam Yadav, and Sangjune Laurence Lee. 2022. Uw-madison gi tract image segmentation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. 2018. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. 2023. Clip-driven universal model for organ segmentation and tumor detection. *arXiv preprint arXiv:2301.00785*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, and 1 others. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. 2018. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 421–429. Springer.

Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. 2021. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Rongguang Ye, Ranmin Wang, Yantong Guo, and Lei Chen. 2022. Sia-unet: A unet with sequence information for gastrointestinal tract segmentation. In *Pacific Rim International Conference on Artificial Intelligence*, pages 316–326. Springer.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

Zheng Zhang, Guanchun Yin, Bo Zhang, Wu Liu, Xiuzhuang Zhou, and Wendong Wang. 2025. A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25940–25949.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867.