PROMPT U-NET: A LEAP TOWARDS CLINICIAN-GUIDED AI FOR CONTEXT-AWARE MEDICAL SEGMENTATION

Paul Machauer[⋆†], *Marco Reisert*[⋆], *Janis Keuper*[†]

[⋆] Division of Medical Physics, Department of Diagnostic and Interventional Radiology, University Medical Center Freiburg, Faculty of Medicine, University of Freiburg, Germany
[†] Offenburg University of Applied Sciences, Offenburg, Germany

ABSTRACT

Automated medical image segmentation has reached high accuracy, yet current models often lack generalizability and clinician control. We introduce ***Prompt U-Net***, a novel convolutional architecture that integrates prompt-based conditioning to address these limitations. The model employs a dual-encoder structure processing both the medical image and a user-provided 2D prompt, enabling rapid adaptation to unseen tasks with limited data. Through a conditioning mechanism and two feedback loops - a self-supervised feedback (SSF) and an interactive feedback (IF) - *Prompt U-Net* supports iterative refinement and ensures volumetric consistency. Experiments on MRI datasets demonstrate superior Dice performance over established baselines with substantially reduced data requirements. These findings highlight prompt-driven convolutional models as lightweight, generalizable, and user-controllable solutions for clinical segmentation.

Index Terms— medical image segmentation, in-context learning, U-Net, generalizability, MRI

1. INTRODUCTION

Medical image segmentation is fundamental to modern diagnostics and treatment planning. Deep learning architectures such as U-Net [1] have achieved remarkable results, yet their static nature limits adaptability and user interaction. Most models function as black boxes, requiring extensive task-specific data while lacking mechanisms to incorporate expert guidance during inference.

Prompt U-Net aims to bridge this gap by transforming segmentation from a static prediction task into an interactive, context-aware process. Inspired by prompting concepts from natural language processing, our model conditions predictions on a user-provided 2D segmentation from a nearby slice. This enables adaptation to new anatomical contexts with minimal out of context training data, fostering true in-context learning rather than memorization.

Our main contributions are:

- A novel dual-encoder U-Net with a conditioning mechanism that fuses multi-scale image and prompt features.
- A training scheme that simulates realistic user interactions and robust prompt perturbations.
- Two feedback loops - self-supervised (SSF) and interactive (IF) - for supporting expert-guided refinement.

The experiments demonstrate that *Prompt U-Net* generalizes effectively across datasets and achieves superior results in a mixture of 2D and 3D tasks with minimal data requirements. This represents a step towards clinician-guided, adaptive AI in medical imaging.

2. RELATED WORK

Our work on *Prompt U-Net* for semi-automatic 3D medical image segmentation intersects the fields of interactive segmentation and in-context learning (ICL). A critical limitation of many current methods is their reliance on pre-defined training distributions, which restricts their adaptability to entirely new anatomical structures. This section reviews relevant works, highlighting this gap which our approach specifically addresses.

In-Context Learning for Medical Segmentation. ICL allows models to perform new tasks based on provided examples without weight updates. The Segment Anything Model (SAM) [2] and its medical adaptation, MedSAM [3], are prominent 2D promptable models. However, their performance is tightly coupled with the domains present in their training data, limiting their zero-shot capability on truly novel structures.

2D Interactive Segmentation. A closer parallel to our work are ICL methods that can reuse a single prompt or a set of prompts for task identification like [4, 5, 6, 7, 8, 9, 10, 11]. *UniverSeg* [4] - trained on 53 datasets - is a leading example, employing a fully convolutional U-Net with "CrossBlocks" to iteratively fuse features from a support support set and a query image. However, *UniverSeg*'s reliance on a pre-defined support set limits its feasibility for real-time, interactive use where a user provides a single prompt on the fly. In contrast,

Prompt U-Net is optimized for efficiency and controllability with a single prompt, making it more suitable for an interactive annotation workflow.

3D Interactive Segmentation. *nnInteractive* [12] is a powerful, general-purpose model trained on an unprecedented scale (120+ datasets) supporting diverse prompts to generate full 3D masks. Its "early prompting" strategy, which incorporates prompts as additional input channels to a 3D U-Net [13], is highly effective. However, its generalizability, while robust across many medical domains, is ultimately bounded by its training distribution. In contrast, our *Prompt U-Net* is designed for a different use case: rapid adaptation from a single example provided by the user within the target volume. This allows our model to segment structures not represented in any pre-training corpus, leveraging the anatomical context of the specific scan.

3. METHOD

3.1. Architecture

The *Prompt U-Net* is based on the U-Net architecture [3] but extends it with a dual-encoder design as shown in fig. 1. As its backbone, the model uses solely standard convolutions [14]. The core components are:

- Image Encoder: Processes the input MRI slice.
- Prompt Encoder: Processes the user-provided prompt (a mask from a nearby slice).
- Conditioning Mechanism: The outputs of both encoders are combined at each layer of the U-Net. This ensures the model leverages both the anatomical context of the query and the specific guidance from the user's prompt at multiple scales.

3.2. Training

To enable efficient training and evaluation of the *Prompt U-Net*, a dedicated data generator was developed. This generator is capable of dynamically producing data points during both training and testing phases by simulating user prompts through spatial offsets in medical image volumes.

For each MRI volume, the generator first performs **slice selection**. A query slice x_k with its corresponding ground truth annotation y_k is randomly chosen, together with a prompt slice x_{k+i} and its annotation y_{k+i} at an offset $i \in [-v, v] \setminus \{0\}$, where v is set to either 7 or 12, depending on the dataset. The combination of query, prompt, and ground truth forms a single data point. Furthermore, for each data point, a random number of segmentations between one and four is selected to increase variability.

To ensure valid prompt–target pairs, two complementary validation strategies are employed. The first, an **intersection check**, verifies that the overlap between annotations is non-empty ($y_k \cap y_{k+i} \neq \emptyset$). The second, a faster alternative based on **label ID matching**, compares the unique annotation identifiers.

After data generation, several **augmentation** steps are applied. First, **synchronized transformations** are applied equally to the query, prompt, and annotation slices. These include spatial scaling and translation of up to $\pm 5\%$, rotations of up to 0.05 radians, and random horizontal or vertical flipping. Next, **prompt-only augmentations** are introduced to simulate imperfect user inputs by randomly corrupting the prompt segmentation y_{k+i} specifically. Up to 30% of the foreground pixels may be set to background. Finally, a variety of **volume crops** are created from the MRI scans prior to slice selection. To improve robustness, 70% of the data points are sampled from cropped volumes and 30% from the original, uncropped volumes.

Normalization is applied to all slices before being used by the model. The segmentation masks y_{k+i} and y_k are binary by definition and are binarized if necessary. The image slices x_{k+i} and x_k undergo a robust min–max normalization defined as:

$$q_{\min} = Q_{l_q}(I), \quad q_{\max} = Q_{u_q}(I), \tag{1}$$

$$I' = \mathrm{clip}(I, q_{\min}, q_{\max}), \tag{2}$$

$$\hat{I} = \frac{I' - q_{\min}}{q_{\max} - q_{\min} + \varepsilon}, \tag{3}$$

where I denotes the input image, Q_{l_q} and Q_{u_q} are the lower and upper quantiles, and ε is a small constant to prevent division by zero.

Models like *UniverSeg* [4] and *mInteractive* [12] were trained on 53 and over 120 datasets respectively, our model was trained on only one dataset: 7 MRI volumes (brain version) or 10 MRI volumes (whole body version). This underscores *Prompt U-Net*'s ability to achieve accurate and generalizable segmentation with substantially reduced data requirements. MRI data is based on the whole body VIBE protocols from [15] and the ground truth labeling is based on [16].

3.3. Inference

For slice-wise inference, the input format follows the same structure as described in the training phase (Section 3.2). To extend the model's applicability to volumetric data, several additional steps are implemented, as shown in Figure 2.

First, the input volume undergoes **preprocessing**, which includes slice-wise min-max normalization and optional cropping to focus on relevant anatomical regions.

Prompt U-Net processes the volume sequentially along the user-specified axis, using the initial prompt slice as the starting point. From this center the segmentation proceeds through all slices in the downward direction, followed by processing of all slices in the upward direction. To maintain segmentation quality across this sequential processing, two complementary feedback mechanisms are employed:

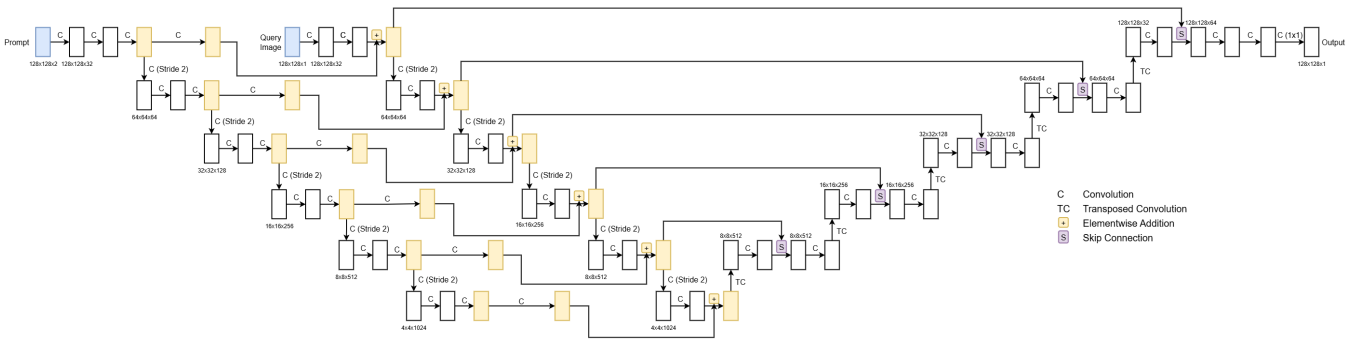

Fig. 1. *Prompt U-Net* Architecture: The white rectangles correspond to the feature maps, consisting of height, width, and number of channels. The blue rectangles represent the inputs to the network, and the yellow ones show the feature maps created by adding the feature maps of both encoders. Convolutions are marked with C and skip connections with a purple S. All convolutions used 3×3 kernels. Encoder filter sizes: 32, 64, 128, 256, 512; decoder: 512, 256, 128, 64, 32; final 1×1 convolution produced the output.

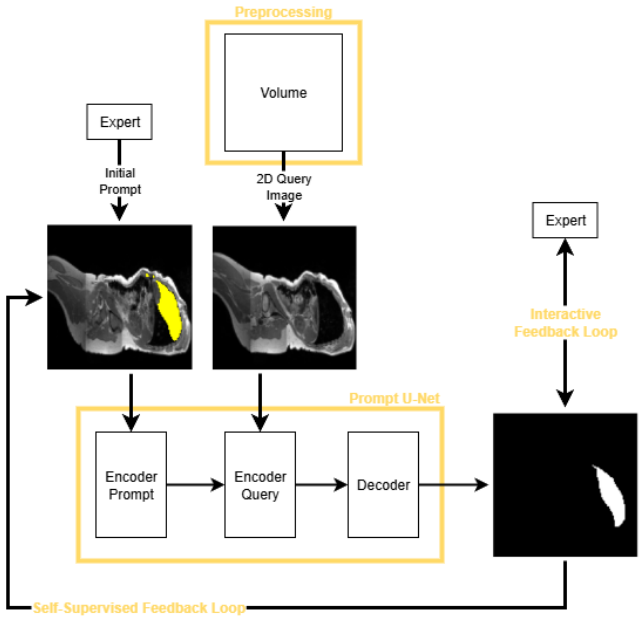

Fig. 2. Sequential slice-wise inference of *Prompt U-Net* on a volumetric scan, conditioned on an initial 2D prompt. The resulting segmentation for one slice serves as the prompt for the next, propagating the segmentation bidirectionally through the volume. This process, supported by self-supervised and interactive feedback loops, ensures robust and consistent 3D segmentation.

- The **Interactive Feedback Loop (IF)** allows users to manually correct predictions during inference. The corrected segmentation then serve as a new prompt for subsequent slices, enabling iterative refinement (Fig. 2).
- The **Self-Supervised Feedback Loop (SSF)** automatically monitors structural consistency between slices. This prevents error propagation when significant anatomical changes occur between consecutive slices. At each iteration i, the structural similarity is computed as:

$$\begin{aligned} s_i &= \mathrm{SSIM}(x_{k+i}, x_k), \\ \Delta s_i &= s_0 - s_i, \\ \text{if } \Delta s_i \geq \tau: \quad y_p &= \mathrm{sign}\big(\min_{b \in \mathcal{B}} b\big). \end{aligned} \tag{4}$$

When the similarity drop Δs_i exceeds threshold τ, the system automatically selects a new prompt segmentation y_p from buffered predictions $\mathcal{B}$. This ensures prompt updates occur when significant anatomical changes are detected between slices.

4. EXPERIMENTS

4.1. 2D Performance

This experiment evaluates the 2D segmentation capability of *Prompt U-Net* using the model variant trained on combined brain and body MRI data. We benchmark our approach against the pre-trained *UniverSeg* model to assess performance in segmenting nearby slices within specified offset ranges using only a single 2D prompt.

All experiments utilize test datasets derived from body MRI scans from [15] (excluding brain MRI) and the labels from [16], as is also the case for the training (3.2). Each subsequent dataset contains data points consisting of query images, prompt slices, and corresponding ground truth annotations. All data points within a test set originate from the same segmentation task and anatomical axis, a constraint imposed by *UniverSeg*'s requirement for consistent prompt sets. Each evaluation dataset comprises 100 query images with 16 additional samples reserved for *UniverSeg*'s prompt set construction.

As shown in Table 1, *Prompt U-Net* achieves a 7.1% higher Dice score than *UniverSeg* for the smaller offset range

Table 1. Comparison of mean Dice coefficients for *Prompt U-Net* (trained on brain and body MRI), *Prompt U-Net* (trained exclusively on brain MRI), and *UniverSeg* at different maximum slice offsets.

Offset	*Prompt U-Net* (combined)	*UniverSeg*	*Prompt U-Net* (brain)
± 5	0.841	0.785	0.775
± 10	0.780	0.787	0.683

(± 5 slices), demonstrating superior performance with a single prompt. The Dice coefficient of 0.841 confirms that our model outperforms established baselines without requiring extensive training datasets or heavily optimized architectures. For larger offset ranges (± 10 slices), *Prompt U-Net* maintains competitive performance (Dice: 0.780). This balance of outstanding low-offset and competitive high-offset performance makes it ideally suited for an interactive process.

4.2. Generalizability

This experiment aims to validate that *Prompt U-Net* does not merely memorize anatomical patterns but genuinely learns to adapt from user-provided contextual instructions.

We employ the variant of *Prompt U-Net* trained exclusively on brain MRI. To assess generalization, we evaluate the model on a body MRI dataset containing segmentation tasks from anatomical regions not present during training. This ensures that the model cannot rely on memorized structural priors. Test data are generated identically to the training procedure (Section 3.2). During each test iteration, a random query image is paired with a simulated user prompt from a nearby slice.

As shown in Table 1, the brain-trained *Prompt U-Net* exhibits only a 7.9% reduction in mean Dice score for the smaller offset range (± 5 slices) compared to the model trained on combined brain and body data. For the larger offset range (± 10 slices), the performance drop increases to 12.4%. These results underscore the model's ability to generalize effectively to data distributions entirely absent from its training set.

4.3. Volumetric Performance

Although *Prompt U-Net* was developed for 2D segmentation, it extends effectively to volumetric data segmentation, aligning with the capabilities of *nnInteractive* which serves as our benchmark. This demonstrates that the 2D training inherently captures 3D anatomical understanding while maintaining prompt adherence, enabling broad application with minimal training.

As shown in Figure 2, we employ self-supervised and interactive feedback loops for sequential slice processing, using the same model from Section 4.1 on full MRI volumes without extracting 2D data points. We conduct two sub-experiments:

1. Performance comparison using only a single initial prompt, relying solely on self-supervised feedback.
2. Interactive scenario incorporating ground-truth prompts when segmentation quality deteriorates.

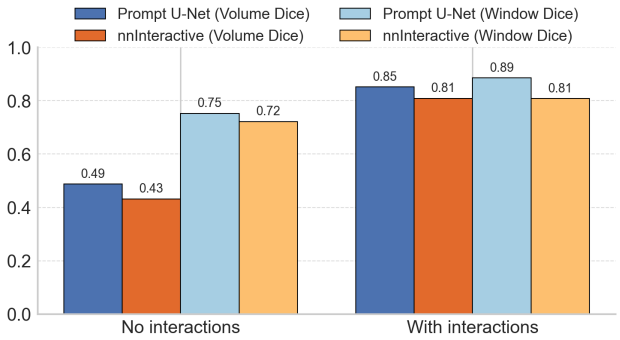

Fig. 3. Volumetric and windowed Dice comparison between *Prompt U-Net* and *nnInteractive* for two scenarios: (1) initial prompt only, and (2) initial prompt plus 14 average interactions. Windowed Dice evaluates performance within ±10 slices of the initial prompt.

As observed in Figure 3 *Prompt U-Net* outperforms *nnInteractive* (Dice: 0.49 vs. 0.43) in the first experiment. The modest absolute scores reflect increasing complexity with distance from the prompt slice, particularly notable given the average of 126 slices per segmentation task. For localized evaluation, a windowed Dice metric shows strong performance for both methods (*Prompt U-Net*: 0.75, *nnInteractive*: 0.72). The self-supervised threshold (0.6) was not optimized and may vary across datasets.

In the interactive experiment, *Prompt U-Net* achieved a mean Dice of 0.85 with an average of 14 additional interactions, demonstrating competitiveness for interactive volumetric segmentation tasks.

5. DISCUSSION

Prompt U-Net demonstrates that a lightweight, convolution-based architecture can achieve strong in-context learning and generalization for medical segmentation. By integrating a 2D prompt, the model effectively adapts to new tasks and anatomical contexts without retraining. In terms of data efficiency, *Prompt U-Net* demonstrates notable performance with a very small training set. While performance is already competitive with larger models, future work could further enhance performance through architectural modernizations. Overall, *Prompt U-Net* offers a practical, user-controllable tool for clinical segmentation workflows.

6. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

7. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[3] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, pp. 654, 2024.

[4] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca, "Universeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21438–21451.

[5] Marianne Rakic, Hallee E Wong, Jose Javier Gonzalez Ortiz, Beth A Cimini, John V Guttag, and Adrian V Dalca, "Tyche: Stochastic in-context learning for medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11159–11173.

[6] Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen, "A simple image segmentation framework via in-context examples," *Advances in Neural Information Processing Systems*, vol. 37, pp. 25095–25119, 2024.

[7] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6830–6839.

[8] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," *arXiv preprint arXiv:2305.13310*, 2023.

[9] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li, "Personalize segment anything model with one shot," *arXiv preprint arXiv:2305.03048*, 2023.

[10] Lingdong Shen, Fangxin Shang, Xiaoshuang Huang, Yehui Yang, Haifeng Huang, and Shiming Xiang, "Segicl: A multimodal in-context learning framework for enhanced segmentation in medical imaging," *arXiv preprint arXiv:2403.16578*, 2024.

[11] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh, "Slime: Segment like me," *arXiv preprint arXiv:2309.03179*, 2023.

[12] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al., "nninteractive: Redefining 3d promptable segmentation," *arXiv preprint arXiv:2503.08373*, 2025.

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[14] Keiron O'shea and Ryan Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[15] Annette Peters, German National Cohort NAKO Consortium, Karin Halina Greiser, Susanne Göttlicher, Wolfgang Ahrens, Maren Albrecht, Fabian Bamberg, Till Bärnighausen, Heiko Becher, Klaus Berger, et al., "Framework and baseline examination of the german national cohort (nako)," *European journal of epidemiology*, vol. 37, no. 10, pp. 1107, 2022.

[16] Robert Graf, Paul Platzek, Evamaria Olga Riedel, Constanze Ramschütz, Sophie Starck, Hendrik K Möller, Matan Atad, Henry Völzke, Robin Bülow, Carsten Oliver Schmidt, et al., "Vibesegmentator: full body mri segmentation for the nako and uk biobank," *European Radiology*, pp. 1–15, 2025.