scientific reports

OPEN

DescriptorMedSAM: language-image fusion with multi-aspect text guidance for medical image segmentation

Wenjie Zhang[1], Liming Luo[2], Mengnan He[2], Jiarui Hai[3] & Jiancheng Ye[1,2]✉

Accurate organ segmentation is essential for clinical tasks such as radiotherapy planning and disease monitoring. Recent foundation models like MedSAM achieve strong results using point or bounding-box prompts but still require manual interaction. We propose DescriptorMedSAM, a lightweight extension of MedSAM that incorporates structured text prompts, ranging from simple organ names to combined shape and location descriptors to enable click-free segmentation. DescriptorMedSAM employs a CLIP text encoder to convert radiology-style descriptors into dense embeddings, which are fused with visual tokens via a cross-attention block and a multi-scale feature extractor. We designed four descriptor types: Name (N), Name + Shape (NS), Name + Location (NL), and Name + Shape + Location (NSL), and evaluated them on the FLARE 2022 dataset under zero-shot and few-shot settings, where organs unseen during training must be segmented with minimal additional data. NSL prompts achieved the highest performance, with a Dice score of 0.9405 under full supervision, a 76.31% zero-shot retention ratio, and a 97.02% retention ratio after fine-tuning with only 50 labeled slices per unseen organ. Adding shape and location cues consistently improved segmentation accuracy, especially for small or morphologically complex structures. We demonstrate that structured language prompts can effectively replace spatial interactions, delivering strong zero-shot performance and rapid few-shot adaptation. By quantifying the role of descriptor, this work lays the groundwork for scalable, prompt-aware segmentation models that generalize across diverse anatomical targets with minimal annotation effort.

Keywords Medical image segmentation, Prompt-aware foundation model, Zero-shot and few-shot learning, Radiology-style anatomical prompts, Multi-scale cross-attention

Accurate segmentation of abdominal organs is critical for tasks such as radiotherapy planning and disease monitoring. Traditional deep learning models, including U-Net[1] and its derivatives (e.g., V-Net[2], PSPNet[3], and TransUNet[4], achieve high Dice scores when trained on thousands of annotated organ masks. However, generating these annotations is both costly and time-consuming for clinicians[5]. To reduce this burden, researchers have explored interactive segmentation methods that require only a few user-provided prompts[6,7]. The Segment Anything Model (SAM)[8] enables high-quality segmentation with minimal input, and MedSAM extends this approach to computed tomography (CT) and magnetic resonance imaging (MRI). Yet, even these spatial prompts, such as clicks on small nodules still demand expert involvement. This limitation has motivated the exploration of even more accessible prompts, including natural language descriptions[9].

Large vision–language models, such as Contrastive Language-Image Pre-training (CLIP)[10], bridge visual and textual domains by aligning an image encoder with a text encoder trained on millions of image–caption pairs[11]. This alignment allows free-form text to guide segmentation without explicit clicks. Recent work has applied similar approaches to medical imaging: FLanS[12] generates free-form text to guide segmentation, and STPNet[13] integrates scale-aware text prompts to enhance performance. However, these studies rarely examine how the granularity of textual prompts influences segmentation quality - a key factor for minimizing annotation effort and ensuring reliable clinical use[14]. Moreover, most prior work evaluates only on organs present in training data, leaving the challenges of zero-shot and few-shot generalization to unseen anatomy largely unaddressed.

[1]Weill Cornell Medicine, Cornell University, New York, NY, USA. [2]Northwestern University, Chicago, IL, USA. [3]Johns Hopkins University, Baltimore, MD, USA. ✉email: jiancheng.ye@u.northwestern.edu

Building on these observations, we propose DescriptorMedSAM, a parameter-efficient extension of MedSAM. The model adds only 1.9 M additional parameters (≈ 1.7% of the base model) and introduces minimal inference overhead while enabling structured language guidance. Our method employs a CLIP text encoder that converts radiology-style descriptors -ranging from simple organ names to detailed prompts such as "Isolate the liver, a large wedge-shaped organ in the right upper quadrant beneath the diaphragm" - into dense embeddings. We systematically study prompt granularity by designing four types of descriptors: Name (N), Name + Shape (NS), Name + Location (NL), and Name + Shape + Location (NSL).

These structured prompts are evaluated under zero-shot and few-shot scenarios, where the model must segment organs unseen during training[15]. In the few-shot setting, fine-tuning with only 50 labeled slices per organ significantly boosts performance, highlighting the potential of structured language to enable rapid adaptation to new anatomical targets. Our contributions are as follows:

1. We introduce a parameter-efficient architecture that integrates linguistic descriptors into MedSAM via a cross-attention mechanism, improving abdominal organ delineation.
2. We develop and evaluate a structured taxonomy of four radiology-style prompts (N, NS, NL, NSL) across 12 abdominal organs from the FLARE-22 dataset[16] under both zero-shot and few-shot settings.

Methods

Dataset

All experiments were conducted on the FLARE 2022 dataset[16], a benchmark comprising 50 contrast-enhanced abdominal CT volumes with expert annotations for 13 organs: liver, spleen, pancreas, stomach, gallbladder, duodenum, esophagus, aorta, inferior vena cava (IVC), left and right kidneys, and left and right adrenal glands. Each volume is provided in 3D NIfTI format. To adapt this data to our 2D segmentation pipeline, we resampled scans to a uniform pixel spacing and stored them as compressed NumPy archives (.npz), resulting in 24,234 fully labeled slices. For consistency across organs, any class with fewer than 2,000 labeled slices was excluded; consequently, the duodenum was omitted, leaving 12 organs for training, validation, and testing. To avoid data leakage arising from the substantial slice-to-slice similarity within each CT volume, the train, validation and test partitions were performed at the volume level rather than the slice level. All slices originating from the same 3D volume were assigned to the same split, ensuring the model is evaluated on volumes unseen during training and mitigating bias from intra-volume redundancy.

Baseline MedSAM

MedSAM processes 2D slices using a frozen ViT-B/16 image encoder pretrained on large-scale natural and medical image datasets. The encoder converts input slices into patch-level visual tokens. A separate prompt encoder embeds user-provided guidance (e.g., points or bounding boxes) into tokens. These visual and prompt tokens are concatenated and passed to a mask decoder, which outputs a probability map for the target structure.

DescriptorMedSAM architecture

The full architecture of DescriptorMedSAM is illustrated in Fig. 1. It retains MedSAM's frozen image encoder and mask decoder while introducing three lightweight modules: (i) Multi-scale feature extractor, (ii) Cross-attention block, and (iii) CLIP-based text-prompt encoder.

First, intermediate features from four layers $\{l_1, l_2, l_3\}$ of the ViT backbone[12] are aggregated and concatenated. A 1 × 1 convolutional neck followed by pixel-shuffling produces a multi-scale feature pyramid, preserving contextual cues while recovering fine spatial details.

Second, radiology-style descriptors are embedded using a CLIP text encoder[10], and the resulting textual tokens are projected to match the dimensionality of visual tokens.

Third, Text embeddings are injected into the visual stream via a multi-head cross-attention mechanism[17]. Let $V \in \mathbb{R}^{N \times d}$ be the visual tokens and $t \in \mathbb{R}^{d_t}$ be the CLIP text embedding. After a linear projection $P_t \in \mathbb{R}^{d_t \times d}$, the block computes.

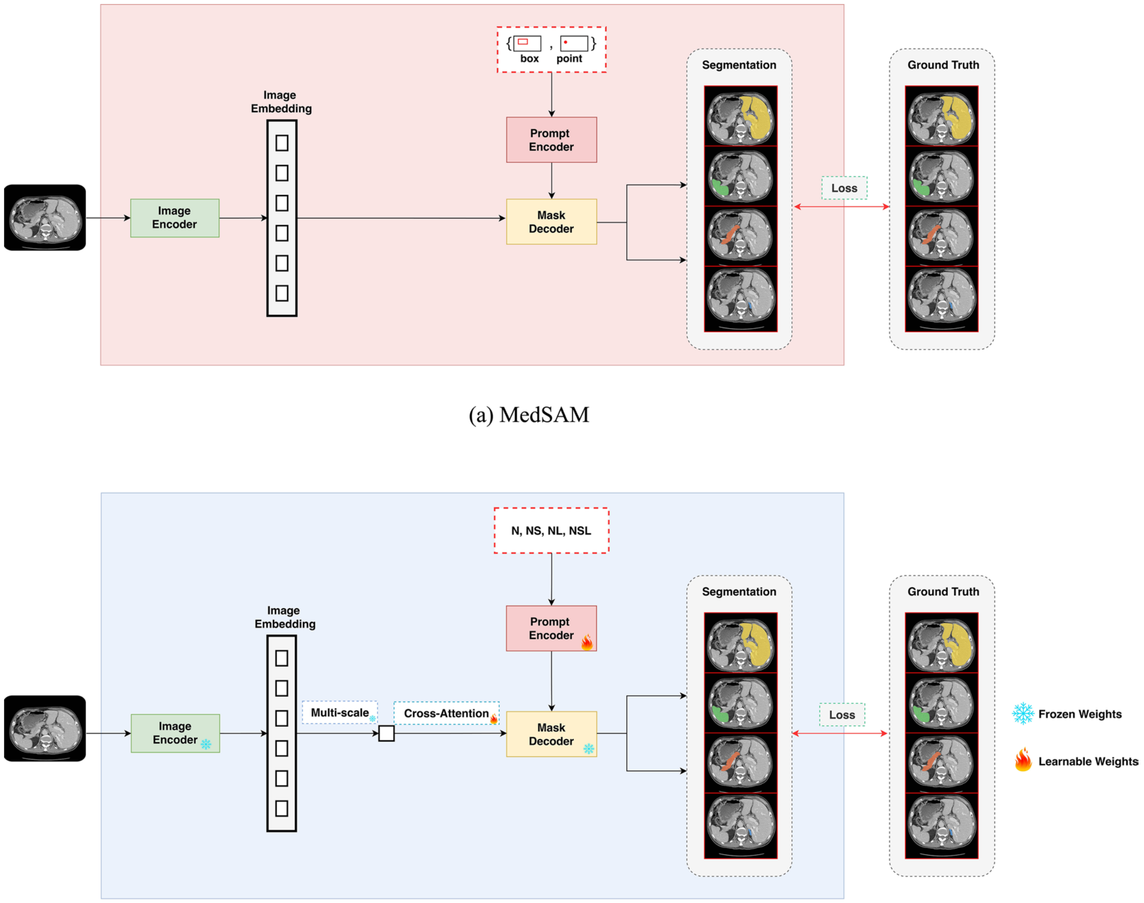
$$\begin{aligned} Q &= VW_Q, \\ K &= (P_t t)W_K, \\ U &= (P_t t)W_V, \\ \tilde{V} &= LN(V + MHA(Q, K, U)) \end{aligned} \tag{1}$$

Where $W_Q,\ W_K,\ W_V \in \mathbb{R}^{d \times d}$ are learned linear maps, $MHA(\cdot)$ is standard multi-head attention, and $LN(\cdot)$ denotes layer normalization with a residual connection. The enhanced token V are concatenated with the multi-scale features and fed into the mask decoder. This modification adds only 1.9 million trainable parameters (≈ 1.7% of total model size), maintaining computational efficiency.

We optimize the decoder with a compound loss that sums dice[18]and binary cross-entropy (BCE) :

$$\mathcal{L} = \lambda_D \mathcal{L}_{Dice} + \lambda_B \mathcal{L}_{BCE} \tag{2}$$

Where (with logits z and ground truth mask $y \in \{0,1\}^{H \times W}$, $\widehat{y} = \sigma(z)$):



(a) MedSAM

(b) DescriptorMedSAM (ours)

Fig. 1. Comparison between (**a**) MedSAM and (**b**) DescriptorMedSAM.

Prompt Category	Generation prompt to GPT-4	Example	Average token count (SD)
N	/	Liver	1.67 (0.85)
NS	Generate one concise radiology-style sentence that describes the typical shape of the {organ}.	Segment the curved, oval spleen hugging the stomach fundus.	10.40 (1.30)
NL	Generate one concise sentence that states the anatomical location of the {organ} using neighboring structures as landmarks.	Mark pancreas body crossing anterior to the aorta.	11.10 (1.06)
NSL	Generate one concise sentence that combines both shape and location information for the {organ}.	Segment the right adrenal gland; note its triangular outline positioned atop the right kidney, posterior to the liver.	16.05 (1.83)

Table 1. Radiological prompt categories, illustrative examples, and GPT-4 generation templates.

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum \widehat{y}y + \epsilon}{\sum \widehat{y} + \sum y + \epsilon} \tag{3}$$

$$\mathcal{L}_{BCE} = -\frac{1}{HW}\sum \left[ylog\widehat{y} + (1-y)log(1-\widehat{y})\right] \tag{4}$$

We set $\lambda_D = \lambda_B = 1$ and $\epsilon = 10^{-6}$ for numerical stability.

Radiological prompt construction

To translate anatomical knowledge into textual prompts, we designed four descriptor prompt categories: N, NS, NL, and NSL, which progressively incorporate name, shape, and location cues. The examples of prompts are as follows:

The N prompt contains only the organ's lowercase name, representing the most minimal user query. For the N prompt category, we did not use GPT-4 generation and duplicated the organ names 20 times. Each of the 12 organ names was duplicated 20 times to match the total prompt count of the other categories. This ensures that all four prompt categories contain 240 prompts for consistent training and evaluation. NS prompts enrich this baseline with morphological cues; a spleen slice might be guided by "Segment the curved, oval spleen hugging the stomach fundus." NL prompts anchor the structure within axial anatomy by citing neighboring organs and directional landmarks, such as "Mark pancreas body crossing anterior to aorta." NSL prompts combine these two aspects into one fluent instruction. For example," segment the right adrenal gland; note its triangular outline positioned atop the right kidney, posterior to the liver."

To generate such sentences, we supply GPT-4[19] with one template per prompt category, substituting the placeholder {organ} with each target-organ name. For every organ, we request 20 variants, resulting in 240 prompts per category (12 organs × 20) and a total of 960 prompts across the four categories. Batched API calls are executed with temperature 0.7, top-p 0.9, and a fixed random seed, guaranteeing that identical calls regenerate the corpus byte-for-byte. The resulting sentences are concise (median ≈ 12 tokens) but provide expressive guidance for every organ.

Experimental protocol

To comprehensively evaluate DescriptorMedSAM under different supervision levels, we adopt three complementary learning protocols: fully supervised, zero-shot, and few-shot segmentation. For full-supervised experiments, the model is trained on all available slices for each organ using the standard FLARE22 training split, following prior MedSAM and medical segmentation literature.

In the zero-shot stage, the model is trained using labeled slices from a randomly selected subset of 8 seen organs and evaluated on the 4 remaining unseen organs. To reduce sampling bias and ensure robustness, this organ-level split is repeated for five independent rounds (Rounds 1–5)[21]. In the Round r, we use $Dice_{ZS}$ to denote the dice on those unseen organs and $Dice_F$ to denote the fully supervised dice. We report the zero-shot retain ratio as

$$RR_{ZS} = \frac{Dice_{ZS}}{Dice_F} \tag{5}$$

In the few-shot stage, following the protocol in SAM Few-shot Finetuning for Anatomical Segmentation in Medical Images[20], we fine-tune the model starting from the zero-shot checkpoint using 50 labeled axial slices per unseen organ. These 50 slices are randomly sampled from all annotated slices of that organ. Because slices from the same 3D CT volume are highly correlated, random sampling may occasionally include multiple slices from a single volume. To mitigate potential bias arising from this limited within-organ diversity, we adopt five independent sampling rounds (Rounds 1–5)[21], each yielding a distinct subset of 50 slices. The final few-shot performance is reported as the mean and 95% confidence interval across these five rounds, thereby substantially reducing sensitivity to specific sampling instances. After the few-shot adaptation, the model is then tested on the remaining slices for those organs. The dice score is $Dice_{FS50}$, and the few-shot retain ratio is computed as

$$RR_{FS} = \frac{Dice_{FS50}}{Dice_F} \tag{6}$$

To ensure robustness and mitigate sampling bias, we repeat the experiment five times (Rounds 1–5) for every prompt type[21]. The same seed is applied across all prompt types—N, NS, NL, and NSL—ensuring that each strategy is evaluated under identical training and test partitions. Within each seen organ, slices are randomly divided into 80% training and 20% validation sets. All hyperparameters are kept constant across all rounds and prompt variants.

Implementation

The proposed framework is implemented on the PyTorch library with one NVIDIA 80G H100 GPU. We adopted the original MedSAM as the base and loaded its official weights. During the training and testing step, the model is trained for 30 epochs on the Flare 2022 dataset with a batch size of 8[16], using the Adam optimizer (learning rate: 0.0001). The baseline models are also trained for 30 epochs on the Flare 2022 dataset with the same hyperparameters. Both SAM and MedSAM receive bounding box prompts during training, following their official configurations. The bounding boxes were derived from the ground-truth masks by computing the tightest enclosing rectangle. These baseline models are also trained for 30 epochs on the Flare 2022 dataset with the same hyperparameters.

Results

Fully-Supervised evaluation

We compared four text-prompt strategies—N, NS, NL, and NSL—against two strong baselines, SAM[8] and MedSAM[22] across 12 abdominal organs. Dice scores are summarized in Table 2.

NSL prompts attained the highest overall performance, achieving an average Dice score of 0.9405 and exceeding the performance of MedSAM and SAM by 1.9 and 5.2% points, respectively. NSL prompts were the top performer for four organs (liver, pancreas, aorta, and left kidney). NS prompts ranked second overall and led in right kidney, spleen, and gallbladder segmentation, indicating that shape information alone significantly enhances delineation. N prompts performed best on IVC and stomach, whereas NL prompts did not lead in any single

Organ	SAM	MedSAM	DescriptorMedSAM-*N*	DescriptorMedSAM-NS	DescriptorMedSAM-NL	DescriptorMedSAM-NSL
Liver	0.9222	0.949	0.9699	0.9819	0.9716	**0.9827**
Right kidney	0.96	0.9449	0.9648	**0.9781**	0.9754	0.9764
Spleen	0.9567	0.957	0.9662	**0.9812**	0.968	0.973
Pancreas	0.8009	0.8567	0.8629	0.8649	0.8749	**0.8903**
Aorta	0.9534	0.9397	0.9557	0.9652	0.9644	**0.9677**
Ivc	0.9272	0.9317	**0.9511**	0.9454	0.947	0.9508
Right adrenal gland	0.7405	**0.8964**	0.8672	0.8701	0.8733	0.8755
Left adrenal gland	0.8084	**0.8851**	0.871	0.8677	0.8717	0.8706
Gallbladder	0.9092	0.905	0.9257	**0.9405**	0.9364	0.9378
Esophagus	0.8601	**0.9264**	0.9066	0.9158	0.9043	0.9137
Stomach	0.8643	0.9193	**0.9713**	0.9685	0.9673	0.9695
Left kidney	0.9539	0.9477	0.9569	0.9773	0.9761	**0.9784**
Average Dice	0.8881	0.9216	0.9308	0.9381	0.9359	**0.9405**

Table 2. Dice scores per organ for SAM, MedSAM, and descriptormedsam with four prompt types.

Model Type	Zero-shot Retain Ratio	Few-shot Retain Ratio
DescriptorMedSAM-N	69.83% (95% CI 59.71%–79.96%)	90.66% (95% CI 79.26%–100.00%)
DescriptorMedSAM-NS	71.17% (95% CI 60.62%–81.72%)	93.94% (95% CI 86.85%–100.00%)
DescriptorMedSAM-NL	72.51% (95% CI 63.99%–81.02%)	93.39% (95% CI 85.62%–100.00%)
DescriptorMedSAM-NSL	**76.31% (95% CI 63.72%–88.90%)**	**97.02% (95% CI 94.20%–99.83%)**

Table 3. Mean zero-shot and 50-shot retain ratios (five-split average). 95% CIs were derived with a two-tailed t-distribution; upper bounds > 100% or lower bounds < 0% were truncated to 100% and 0%, respectively.

organ but consistently delivered competitive results. Interestingly, MedSAM outperformed DescriptorMedSAM on three small, low-contrast organs (left adrenal gland, right adrenal gland, and esophagus), suggesting that such structures remain more sensitive to noise in textual prompts. Overall, NSL prompts emerged as the most reliable choice, improving segmentation accuracy across all organs without degrading performance for any specific case.

Zero- & Few-Shot adaptation

Table 3 summarizes the mean zero-shot and few-shot retention ratios across five experimental rounds.

Adding textual detail consistently improved zero-shot generalization. Compared to N prompts, shape prompts (NS) improved retention by 1.3% points, location prompts (NL) by 2.6 points, and combined prompts (NSL) by 6.5 points. Without additional labels, NSL prompts already preserved 76.31% of fully supervised Dice scores, narrowing the performance gap to less than 25%.

With few-shot fine-tuning using only 50 labeled slices per unseen organ, all prompt types achieved substantial gains. NSL prompts reached 97.02% of fully supervised performance, outperforming NS and NL prompts (~ 94%) and N prompts (~ 91%). These findings indicate that morphological and spatial cues are complementary, with NSL prompts enabling the fastest and highest adaptation to unseen organs.

Confidence intervals for all prompt types remained within ± 15% points, indicating stable performance across experimental rounds. Although some intervals overlap, the NSL prompt distribution consistently lies higher, reinforcing its superior generalization in both zero-shot and few-shot scenarios.

Figure 2 provides a qualitative comparison of DescriptorMedSAM's outputs under different training regimes. The figure shows that the fully supervised model closely replicates the ground-truth masks across all organs. In the zero-shot setting, DescriptorMedSAM with NSL prompts already covers some of the large, high-contrast organs, such as the stomach and left kidney; however, it under-segments smaller or morphologically complex structures, including the pancreas and both adrenal glands. After fine-tuning with just 50 labelled slices per unseen organ, the red contours for these challenging organs tighten markedly, and the visual results approach those of the fully supervised model. This qualitative improvement confirms that a small annotation budget can substantially narrow the performance gap to full supervision.

Overall, NSL provides the strongest zero-shot safety net and the steepest few-shot learning curve once a handful of labelled slices becomes available.

Discussion

Recent advances in medical image segmentation have largely centered on interactive spatial prompting, where models such as SAM[8] and MedSAM[22] rely on user-supplied points or bounding boxes to localize target regions. Variants like ProtoSAM[23] further streamline this interaction. Although these approaches—built on a shared encoder - decoder architecture, delivering state-of-the-art accuracy on organs encountered during training, they share a critical limitation: every new case still demands additional human input. This extra annotation workload

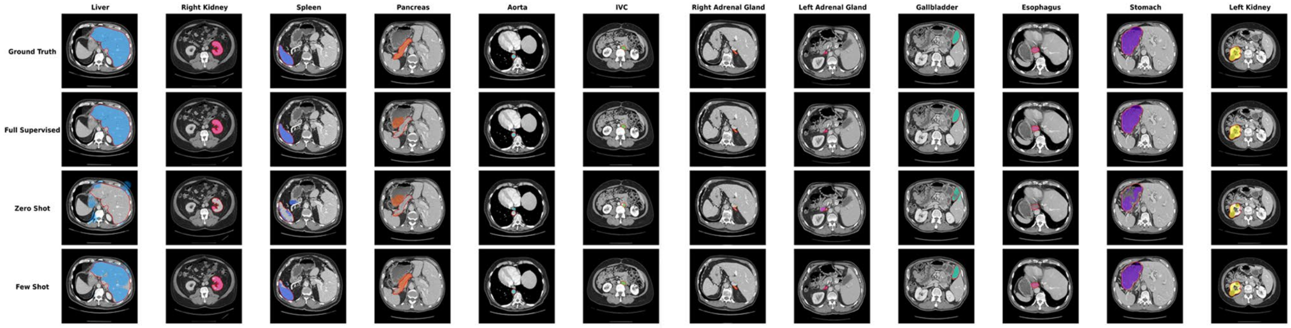


Fig. 2. Qualitative segmentation results for twelve abdominal organs. Columns list the organs (Liver to Left Kidney); rows show, top to bottom: Ground Truth masks, Fully-Supervised DescriptorMedSAM (NSL), Zero-Shot DescriptorMedSAM (NSL), and Few-Shot DescriptorMedSAM after 50 labelled slices per unseen organ. Ground-truth regions are filled with organ-specific colors; predicted boundaries are overlaid in red.

hinders scalability and complicates deployment in time-sensitive clinical settings. To break this dependency on clicks, researchers have begun exploring language-based prompting to guide the segmentation. FLanS[12] employs large language models to generate free-form descriptions, STPNet[13] introduces scale-aware text prompts for lesion segmentation, and TGAM fuses anatomical sentences with visual features[24]. These studies demonstrate that natural-language guidance can effectively replace spatial prompts and integrate seamlessly with radiology reporting. However, they share several limitations: most treat the text prompt as a black-box input without examining its semantic granularity, and they primarily evaluate on anatomies already encountered during training.

To overcome these problems, we introduce DescriptorMedSAM—a descriptor-guided extension of MedSAM. It maintains MedSAM's image encoder and mask decoder in a frozen state to preserve speed and memory, while injecting language via a parameter-To unpack the role of linguistic detail, we devise a four-level prompt taxonomy—N, NS, NL, and NSL—that allows us to measure how each semantic aspect influences performance. Finally, we established a unified zero-shot and few-shot protocol on FLARE 2022, utilizing five random organ splits, which enabled rigorous tests on anatomy never encountered during training and rapid adaptation with just 50 labelled slices per unseen organ. In summary, DescriptorMedSAM captures fine-grained textual information, generalizes effectively to unseen anatomy, and maintains the efficiency required for real-time clinical deployment.

Empirically, finer semantics consistently help: NSL improves the zero-shot retain ratio by 6.5% points over N and reaches 97.02% of fully supervised performance after 50 slices per organ of fine-tuning, with results stable across splits (95% CIs within ± 15 pp). Qualitative results echo these trends: zero-shot predictions already cover large organs well but miss small or intricate structures, whereas few-shot fine-tuning sharpens the boundaries of those challenging regions, bringing them close to fully supervised quality. Although the four descriptor variants (N, NS, NL, NSL) show only modest differences under the fully supervised setting, this trend is reasonable because all variants rely on the same complete set of annotations and share the same MedSAM backbone. Additionally, replacing click-based prompts with text descriptors does not negatively affect accuracy, and the performance of different prompt types naturally converges to similar Dice levels. In contrast, the effect of semantic granularity becomes more pronounced in the zero-shot and few-shot settings, where the model must generalize to organs not seen during training. The richer descriptors (NSL) consistently achieve higher retain ratios and more stable performance across splits, demonstrating that semantic detail primarily benefits generalization rather than fully supervised learning. This contrast highlights that DescriptorMedSAM's improvements lie in its ability to leverage textual semantics to enhance cross-organ generalization, rather than to boost already-saturated fully supervised performance.

These findings clarify that DescriptorMedSAM retains the click-free convenience of language-guided segmentation, while quantifying the impact of prompt semantic granularity and delivering substantial zero-shot and few-shot improvements on previously unseen organs. Beyond quantitative improvements, language-guided segmentation also carries practical advantages in clinical workflows. Unlike point- or box-based prompting, which requires precise manual interaction and introduces inter-operator variability, radiologists and surgeons naturally describe anatomy using text, such as "the superior pole of the left kidney" or "the pancreatic tail adjacent to the spleen." A text-driven interface therefore enables zero-click or hands-free operation, reduces interaction burden in high-volume reading settings, and improves reproducibility by eliminating variations in mouse-based prompt placement. Moreover, structured text can encode subtle clinical attributes that are difficult to express through bounding boxes, making it particularly valuable for irregular or elongated organs such as the pancreas or adrenal glands. Language-guided segmentation thus aligns more closely with real clinical communication patterns and can integrate seamlessly with AI-assisted reporting, surgical navigation, and telemedicine systems.

While our four-level taxonomy (N, NS, NL, NSL) captures three major semantic dimensions—name, shape, and coarse location—it does not exhaust the full spectrum of clinically meaningful descriptors. In practice, radiologists often reference additional attributes such as organ size, relative intensity ("hypodense compared with the liver"), adjacency ("anterior to the spine"), or sub-regional characteristics ("pancreatic tail", "upper pole of the kidney"). Incorporating such fine-grained or context-dependent details may further enhance generalization, although their benefits likely depend on the anatomical complexity and the clarity of the underlying imaging features. Determining how much semantic information is beneficial and at what point additional detail yields diminishing or no further improvements remains an open question. Exploring these limits would require more diverse datasets and clinically curated textual annotations, which we identify as an important direction for future work. Our current study provides an initial step by quantifying how structured descriptor richness affects segmentation performance, but a broader investigation into the full space of clinical semantics represents a promising extension of this framework.

By systematically quantifying the effect of semantic granularity, DescriptorMedSAM is the first to show that combining shape and location cues provides distinctive benefits for delineating organ boundaries. At the same time, it delivers strong zero-shot performance from structured text alone and rapidly approaches fully supervised accuracy with just a handful of labeled slices. This capability can shorten model-development cycles from weeks to days and substantially ease the burden on clinicians.

Limitation

This study has several limitations. First, although the labeled subset of FLARE2022 provides 24k axial slices, these originate from only 50 abdominal CT volumes. As slices within the same volume are highly correlated, the effective diversity of the training set is limited. This is an inherent limitation of slice-based training and may cause the model to over-represent volume-specific appearance patterns. While our volume-level split prevents direct leakage across folds, broader multi-centre datasets or the inclusion of the unlabeled FLARE2022 volumes would further strengthen the generalizability of DescriptorMedSAM. Future work should validate the approach on larger, multi-institutional datasets and across different imaging modalities[25]. Second, our prompt design focused on four fixed categories (N, NS, NL, NSL), which may not capture nuances such as pathology-specific descriptors, temporal context, or radiologist shorthand. Expanding to adaptive or pathology-aware prompts could further enhance performance and robustness. Finally, while our architecture is lightweight, real-time clinical deployment will require additional work to optimize inference speed, integrate into Picture Archiving and Communication System (PACS), and evaluate usability with radiologists in prospective workflows.

Conclusion

This study presents DescriptorMedSAM, an extension of MedSAM that integrates structured descriptor prompts into medical image segmentation via a cross-attention mechanism and multi-scale feature fusion. By systematically varying prompt granularity—from simple organ names to combined shape-location descriptors—we demonstrate that richer textual guidance significantly enhances both zero-shot and few-shot segmentation performance. By quantifying the impact of prompt semantics and demonstrating strong generalization to unseen organs, DescriptorMedSAM establishes a foundation for prompt-aware medical segmentation models. Future work should extend this framework to larger, multi-institutional datasets, additional imaging modalities, and adaptive prompt generation strategies, further advancing the clinical applicability of language-guided segmentation.

Data availability

FLARE22 are available at https://flare22.grand-challenge.org.

Code availability

The relevant code are available at: https://github.com/Wenj1eee/Prompt-Dimensions-of-MedSAM.

Received: 7 October 2025; Accepted: 22 December 2025
Published online: 12 January 2026

References

1. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. (Springer, 2015).
2. Milletari, F., Navab, N. & Ahmadi, S. A. *V-net: Fully convolutional neural networks for volumetric medical image segmentation*. in *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016. Ieee. (2016).
3. Zhao, H. et al. *Pyramid scene parsing network*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017).
4. Chen, J. et al. *Transunet: Transformers make strong encoders for medical image segmentation*. arXiv preprint arXiv:2102.04306, (2021).
5. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image. Anal.* **42**, 60–88 (2017).
6. Sofiiuk, K., Petrov, I. A. & Konushin, A. *Reviving iterative training with mask guidance for interactive segmentation*. in *2022 IEEE international conference on image processing (ICIP)*. IEEE (2022).
7. Chen, X. et al. *Focalclick: Towards practical interactive image segmentation*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2022).
8. Kirillov, A. et al. *Segment anything*. in *Proceedings of the IEEE/CVF international conference on computer vision*. (2023).
9. Liu, S. et al. Cross-modal progressive comprehension for referring segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (9), 4761–4775 (2021).
10. Radford, A. et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. PmLR. (2021).
11. Zhang, Z. et al. *Echo-Vision-FM: A Pre-training and Fine-tuning Framework for Echocardiogram Videos Vision Foundation Model.* medRxiv, 2024.10.09.24315195. (2024).
12. Da, L. et al. *Segment as You Wish–Free-Form Language-Based Segmentation for Medical Images.* arXiv preprint arXiv:2410.12831, (2024).

13. Shan, D. et al. *STPNet: Scale-aware Text Prompt Network for Medical Image Segmentation* (IEEE Transactions on Image Processing, 2025).
14. Ye, J. et al. *DeepSeek in Healthcare: A Survey of Capabilities, Risks, and Clinical Applications of Open-Source Large Language Models*. arXiv preprint arXiv:2506.01257, (2025).
15. Pourpanah, F. et al. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (4), 4051–4070 (2022).
16. Ma, J. et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. *Lancet Digit. Health.* **6** (11), e815–e826 (2024).
17. Vaswani, A. et al. *Attention is all you need. Adv. Neural. Inf. Process. Syst.* **30**. (2017).
18. Sudre, C. H. et al. *Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations*. in *International Workshop on Deep Learning in Medical Image Analysis*. Springer. (2017).
19. Achiam, J. et al. *Gpt-4 technical report*. arXiv preprint arXiv:2303.08774, (2023).
20. Xie, W. et al. *Sam fewshot finetuning for anatomical segmentation in medical images*. in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. (2024).
21. An, D. et al. *Sli2Vol+: Segmenting 3D medical images based on an object estimation guided correspondence flow network*. in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. (2025).
22. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15** (1), 654 (2024).
23. Ayzenberg, L., Giryes, R. & Greenspan, H. *Protosam: One-shot medical image segmentation with foundational models*. arXiv preprint arXiv:2407.07042, (2024).
24. Rahman, M. M. et al. *Text-Assisted Vision Model for Medical Image Segmentation* (IEEE Journal of Biomedical and Health Informatics, 2025).
25. Ye, J. et al. *Multimodal data hybrid fusion and natural language processing for clinical prediction models*. AMIA Summits on Translational Science Proceedings, **2024**, p. 191. (2024).

Author contributions

JY designed the study. WZ, JH, LL, and JY contributed to data analyses. WZ and JY contributed to the writing of the manuscript. MH contributed to data management. All authors read and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.