TuneVLSeg: Prompt Tuning Benchmark for Vision-Language Segmentation Models

Rabin Adhikari, Safal Thapaliya, Manish Dhakal, and Bishesh Khanal

Nepal Applied Mathematics and Informatics Institute for research (NAAMII), Nepal
{rabin.adhikari,safal.thapaliya,manish.dhakal,bishesh.khanal}@naamii.org.np

Abstract. Vision-Language Models (VLMs) have shown impressive performance in vision tasks, but adapting them to new domains often requires expensive fine-tuning. Prompt tuning techniques, including textual, visual, and multimodal prompting, offer efficient alternatives by leveraging learnable prompts. However, their application to Vision-Language Segmentation Models (VLSMs) and evaluation under significant domain shifts remain unexplored. This work presents an open-source benchmarking framework, *TuneVLSeg*, to integrate various unimodal and multimodal prompt tuning techniques into VLSMs, making prompt tuning usable for downstream segmentation datasets with any number of classes. *TuneVLSeg* includes 6 prompt tuning strategies on various prompt depths used in 2 VLSMs totaling of 8 different combinations. We test various prompt tuning on 8 diverse medical datasets, including 3 radiology datasets (breast tumor, echocardiograph, chest X-ray pathologies) and 5 non-radiology datasets (polyp, ulcer, skin cancer), and two natural domain segmentation datasets. Our study found that textual prompt tuning struggles under significant domain shifts, from natural-domain images to medical data. Furthermore, visual prompt tuning, with fewer hyperparameters than multimodal prompt tuning, often achieves performance competitive to multimodal approaches, making it a valuable first attempt. Our work advances the understanding and applicability of different prompt-tuning techniques for robust domain-specific segmentation. The source code is available at `https://github.com/naamiinepal/tunevlseg`.

Keywords: Prompt Tuning · Vision-Language Segmentation Models · Medical Image Segmentation

1 Introduction

Segmenting the anatomical and pathological structures in medical images is crucial for computer-aided diagnosis, prognosis, and surgery planning. Recent deep-learning-based segmentation models have shown excellent performance on curated datasets, but lack generalization across anatomies and image modalities as they are typically trained on a limited set of anatomies and modalities or fine-tuned using pretrained weights from models trained on natural images

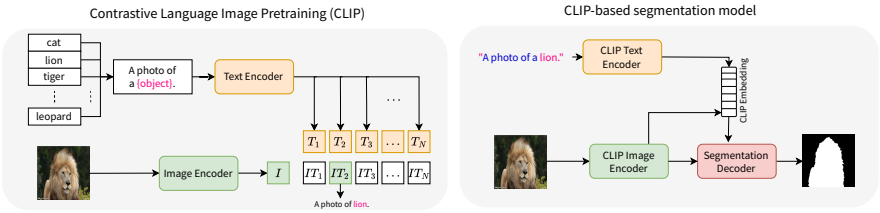


Fig. 1: Manual prompting in CLIP and CLIP-based segmentation models.

like ImageNet [10]. Recent advancements in foundational vision language models (VLMs) [18, 28, 37, 43, 55, 58], which leverage the image and text data, have gained significant attention from the research community due to their generalization capabilities across new dataset and vision tasks. These models have been adopted for segmentation tasks in natural images [32, 38, 45, 63], and show excellent generalization for segmentation as well. These vision-language segmentation models (VLSMs) use the pretrained encoders of the VLMs with an added segmentation decoder. The surprising generalization ability of VLSMs stems from the language supervision provided by text inputs, known as *prompts*. During inference, prompts like "`a photo of a {object}`" supply auxiliary information along with image embeddings to identify the target class, and high-quality prompts are crucial for enhancing VLSM's performance [21, 36].

Extending VLSMs to medical image segmentation tasks presents challenges, particularly in designing effective prompts for VLSMs pretrained on natural images [36]. This often results in suboptimal performance for medical image segmentation, thus requiring fine-tuning on medical datasets [13, 36, 42, 56, 59]. Given the massive scale of these models and the scarcity of large labeled medical datasets, fine-tuning VLSMs for medical datasets is often infeasible. One of the efforts to extend these models to new domains with fewer data and computational requirements is prompt tuning. It is robust to noisy labels [49], making this strategy favorable to extend VLMs pretrained on natural images to medical datasets. The prompt tuning strategies (also known as *context learners*) can be extended to VLSMs as they share a similar architecture with VLMs, which are usually composed of text and vision encoders, with an added segmentation decoder (see Fig. 1).

Prompt tuning addresses the challenge of hand-engineering prompts by introducing learnable prompts (also called *context vectors*), thus adapting VLMs to new datasets by optimizing only these context vectors [62]. Extending prompt tuning to the segmentation task has its caveats. As illustrated in Fig. 2, the context vectors can be introduced at text [61, 62], vision [19], or both inputs [23, 53]. Additionally, these vectors can be injected at different depths at the encoders, and with an added decoder for segmentation tasks, these challenges are further escalated. However, most studies on prompt tuning have focused primarily on image classification, limiting insights on the prospect of prompt tuning on downstream image segmentation tasks.

In this work, we propose an open-source benchmark framework, *TuneVLSeg*, incorporating unimodal and multimodal prompt tuning methods for a principled way of evaluating prompt tuning for different class-agnostic VLSMs. We study the effects of adding the context vectors at multiple depths at both image and text encoders. We also evaluate the performance of the 6 prompt tuning methods on adapting 2 pretrained VLSMs to 2 natural and 8 medical segmentation datasets. Our benchmark framework can be extended to adapting other class-agnostic VLSMs using new prompt-tuning methods to other segmentation datasets with little to no effort.

Our major contributions encompass the following points:

- A principled way of evaluating different unimodal and multimodal prompt-tuning strategies for segmentation tasks.
- A modifiable and reusable benchmark framework, *TuneVLSeg*, leveraging pretrained VLSMs to fine-tune them to target segmentation tasks with prompt-tuning.
- Comparing the effectiveness of different unimodal and multimodal prompt-tuning in natural and medical segmentation datasets across multiple radiology and non-radiology images.

2 Related Work

2.1 Vision-Language Models

Vision-Language Models (VLMs) incorporate representation modules to connect image and text features, enabling their use in various vision-language tasks such as visual question-answering (VQA), image segmentation, image-text retrieval, object detection, and phrase grounding. Bordes *et al*. [6] have categorized VLM training into four distinct classes. *Contrastive VLMs* [18, 28, 37, 54] learn to project text and vision features onto the same embedding space to establish equivalence among the modalities. *VLMs with masking objectives* [8, 24, 43] are trained to guide models to predict missing image patches or text tokens for establishing similarity between both modalities. *Generative models* [39, 52] are optimized for generating images using the text representation and vice-versa. *VLMs with pretrained backbone* [44, 55, 65] train a mapping network between pretrained encoders, thus, requiring fewer compute.

2.2 Vision-Language Segmentation Models

VLMs' capacity to capture multimodal information has led to their widespread use as a backbone for segmentation tasks. Vision-Language Segmentation Models (VLSMs) incorporate a vision-language decoder on top of the pretrained VLM encoders to capture information from both text and vision branches. DenseCLIP [38] employs a vision-language decoder atop CLIP encoders, utilizing pixel-text score maps to guide the learning of dense prediction models. Similarly, CLIPSeg [32] and CRIS [45] provide zero-shot segmentations by predicting pixel-level

activations for the given text or image query. ZegCLIP [63] introduces tuned prompts and associates image information with text encodings before patch-text contrasting, aiming to reduce overfitting of seen classes in both inductive and transductive zero-shot settings.

2.3 Prompt Tuning

Prompts play a crucial role in downstream datasets [21, 51]. However, identifying the optimal prompt for a given downstream task can be burdensome and sometimes requires prior knowledge. Therefore, a more effective approach is automatically learning these prompts for specific downstream tasks instead of providing manual prompts. This paradigm of learning a task-specific prompt is known as prompt tuning, which was first used in NLP [20, 26, 27, 29, 30, 41, 60]. It was subsequently adapted for the text branch of VLMs [61, 62, 64] and vision-only models [19, 46, 47, 57]. Finally, multimodal prompt tuning strategies [23, 53] were applied to both text and vision encoders of VLMs, demonstrating superior performance compared to unimodal approaches.

Although prompt tuning strategies have been extensively studied for VLMs like CLIP [37], few studies have focused on applying these techniques to VLSMs. DenseCLIP [38] and ZegCLIP [63] introduced CoOp [62] and VPT [19] for VLSMs, respectively. However, the main drawback of these models is that changes in the number of classes in the segmentation task alter the number of channels in their weights. This renders their pretrained weights unusable for downstream tasks unless the desired class(es) were present during pretraining.

3 Revisiting CLIP

3.1 Text Encoding

CLIP's text encoder tokenizes the provided text with some special tokens, generates their corresponding word embeddings, and adds the position encoding before passing them to the first transformer layer, $W_0 = [w_0^1, w_0^2, \cdots, w_0^N] \in \mathbb{R}^{N \times H_l}$. Each W_i is passed to the $(i+1)^{th}$ transformer layer $\mathcal{L}_{i+1}$, outputting W_{i+1}. Mathematically,

$$W_{i+1} = \mathcal{L}_{i+1}(W_i) \qquad i = 0, 1, \cdots, K_l - 1 \tag{1}$$

Here, K_l denotes the number of transformer layers for the text encoder. Furthermore, the sentence embedding is extracted from the output the ultimate layer (W_K) as the last vector in the sequence (corresponding to the End of Sentence (EOS) token), *i.e.*, $W_K^N \in \mathbb{R}^{H_l}$. The sentence embedding in the text space is projected into the vision-language space using a learnable linear projection matrix, $\mathcal{P}_l \in \mathbb{R}^{H_l \times H_{vl}}$. Mathematically,

$$z_l = \mathcal{P}_l^T W_K^N \qquad z_l \in \mathbb{R}^{H_{vl}} \tag{2}$$

3.2 Image Encoding

CLIP's image encoder splits the image into multiple patches of predefined sizes. It projects them linearly to obtain token embeddings and adds the position encoding to them before passing to the first ViT layer, *i.e.*, $E_0 = [e_0^1, e_0^2, \cdots, e_0^N] \in \mathbb{R}^{N \times H_v}$. Additionally, a class (CLS) token embedding ($c_i \in \mathbb{R}^{H_v}$) is concatenated to the token embedding (E_i) before passing to the transformer layer $\mathcal{V}_{i+1}$, outputting $[c_{i+1}, E_{i+1}]$. Mathematically,

$$[c_{i+1}, E_{i+1}] = \mathcal{V}_{i+1}([c_i, E_i]) \qquad i = 0, 1, \cdots, K_v - 1 \tag{3}$$

Here, K_v denotes the number of transformer layers for the image encoder. The class (CLS) embedding from the last layer (c_K) corresponds to the global embedding (or image-level embedding), which is in the image space. The embedding is projected into the vision-language space using a learnable linear projection matrix, $\mathcal{P}_v \in \mathbb{R}^{H_t \times H_{vl}}$. Mathematically,

$$z_v = \mathcal{P}_v^T c_K \qquad z_v \in \mathbb{R}^{H_{vl}} \tag{4}$$

4 Revisiting Prompt Tuning

4.1 Textual Prompt Tuning

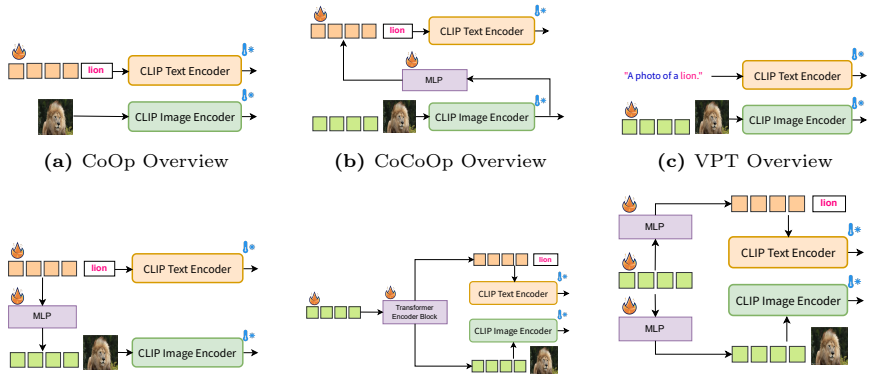
To learn language-specific prompts, we assign B learnable tokens to the first transformer layer for CLIP's text encoder, as $P_0 = [p_0^1, p_0^2, \cdots, p_0^B] \in \mathbb{R}^{B \times H_l}$. The input for the first transformer layer of the text branch is the concatenation of the learnable prompts and provided fixed text prompt, *i.e.*, $[P_0, W_0]$ (see Fig. 2a). Also, these learnable prompts can be extended to multiple layers of the text encoder, such that J is the network depth up to which new prompts are learned for each layer. This deep prompting technique helps align the intermediate layers' outputs with the first layer's output. Mathematically,

$$[\_, W_{i+1}] = \mathcal{L}_{i+1}([P_i, W_i]) \qquad i = 0, 1, \cdots, J - 1 \tag{5}$$

Up to J^{th} transformer layer, the outputs corresponding to learnable prompts P are discarded, and a new set of learnable prompts is injected for each layer. Whereas, after the J^{th} layer ($\mathcal{L}_J$), the computation is carried on similar to the one described in subsection 3.1 as if the prompts are identical to the word embeddings. Mathematically,

$$[P_{i+1}, W_{i+1}] = \mathcal{L}_{i+1}([P_i, W_i]) \qquad i = J, J + 1, \cdots, N_l - 1 \tag{6}$$

When the depth of the network to which the prompt is injected is unity, *i.e.*, $J = 1$, it subsides to CoOp [62]. CoCoOp [61] extended CoOp by adding the projection of CLIP's image-level embeddings to learnable prompts. It makes the textual prompts image-dependent compared to the instance-agnostic task-dependent prompts of CoOp.



(a) CoOp Overview **(b)** CoCoOp Overview **(c)** VPT Overview

(d) Maple Overview **(e)** Shared Attention Overview **(f)** Shared Separate Overview

Fig. 2: Overview of various prompt tuning methods. In the first row, there are unimodal prompt tuning methods and the second row shows the multimodal prompt tuning methods. The prompting for only the first layer is shown here, the same concept is applicable when the prompt tuning is done for multiple transformer blocks.

4.2 Visual Prompt Tuning

Similar to textual prompt tuning, we assign B learnable tokens to the first transformer layer for CLIP's vision encoder, as $\tilde{P}_0 \in \mathbb{R}^{B \times H_v}$ (see Fig. 2c). The input for the first transformer layer of ViT is the concatenation of the CLS token, fixed image token embeddings, and the learnable prompts, *i.e.*, $[c_0, E_0, \tilde{P}_0]$. Mathematically,

$$[c_{i+1}, E_{i+1}, \_] = \mathcal{V}_{i+1}([c_i, E_i, \tilde{P}_i]) \qquad i = 0, 1, \cdots, J-1 \tag{7}$$

Similar to textual prompt tuning, up to J^{th} transformer layer, the outputs corresponding to learnable prompts $\tilde{P}$ are discarded, and a new set of learnable prompts is injected for each layer. Whereas, after the J^{th} layer ($\mathcal{V}_J$), the computation is carried on similar to the one described in subsection 3.2 as if the prompts are identical to the token embeddings. Mathematically,

$$[c_{i+1}, E_{i+1}, \tilde{P}_{i+1}] = \mathcal{V}_{i+1}([c_i, E_i, \tilde{P}_i]) \qquad i = J, J+1, \cdots, N_v - 1 \tag{8}$$

4.3 Multimodal Prompt Tuning

Multimodal approaches to prompt tuning were devised to adjust the output for both the encoders. Similar to unimodal prompt tuning, we assign B tokens to each transformer layer i till the prompt depth J for CLIP's text and vision encoders, as $P_i \in \mathbb{R}^{B \times H_l}$ and $\tilde{P}_i \in \mathbb{R}^{B \times H_v}$, respectively.

A straightforward way to implement the multimodal prompt tuning is to learn both textual (P_i) and visual ($\tilde{P}_i$) prompts independently in the same training schedule. However, studies have shown that such a naive training strategy

leads to sub-optimal performance due to the lack of interaction between the vision and language branches [23, 53]. The interaction between the two modalities can be achieved by introducing unified prompts for each transformer layer i, denoted by $\hat{P}_i \in \mathbb{R}^{B \times H_u}$. Here, H_u, a hyperparameter, is the dimension of the unified prompt. We obtain visual and textual prompts from the corresponding unified prompts as follows.

$$[P_i, \tilde{P}_i] = \mathcal{U}_i(\hat{P}_i) \tag{9}$$

Here, $\mathcal{U}_i$ is a learnable function that transforms unified prompts $\hat{P}_i$ for each transformer layer i into corresponding textual P_i and visual $\tilde{P}_i$ prompts. Some architectures of functions $\mathcal{U}_i$ are self-attention mechanisms, multi-layer perception, or a mixture of both. For this work, we use a transformer block for each $\mathcal{U}_i$ as one of the methods of multimodal prompt tuning, named *Shared Attention* (see Fig. 2e).

One specialized case of these transformations is when the transformations are split into separate components of text and vision transformations, viz. $\mathcal{U}_i^l$ and $\mathcal{U}_i^v$ which transform the unified prompts to the corresponding text and vision spaces. For example, the transformation function is a perceptron (an affine transformation) — and optionally followed by an activation function. Mathematically,

$$\mathcal{U}_i(\hat{P}_i) = [\mathcal{U}_i^l(\hat{P}_i), \mathcal{U}_i^v(\hat{P}_i)] \tag{10}$$

Equivalently,

$$P_i = \mathcal{U}_i^l(\hat{P}_i) \qquad \tilde{P}_i = \mathcal{U}_i^v(\hat{P}_i) \tag{11}$$

For these separate transformations, we use a linear layer followed by layer normalization [4] to project the unified prompts to textual and visual dimensions. We call the method *Shared Separate* (see Fig. 2f). A more specialized case of these separable transforms is when one of the transforms, $\mathcal{U}_i^l(\hat{P}_i)$ or $\mathcal{U}_i^v(\hat{P}_i)$, is identity. In that case, the unified prompts are directly initialized to the space for which the transformation is identity. Khattak *et al*. [23] does the same; they initialize the unified prompts in the text space (see Fig. 2d).

5 Method

Fig. 3 shows the general architecture of our proposed framework. For multimodal prompt tuning, the learnable prompts are injected in both the encoders whereas for unimodal prompt tuning, the prompts are injected in the encoder of the respective modality. Also, since all the parameters of the pretrained models are frozen, we train and store only the learnable contexts for each downstream task, thus saving computational resources and storage.

In our study, we have used two VLSMs that are pretrained on natural images viz. CLIPSeg [32] and CRIS [45] for prompt tuning. A key difference between CLIPSeg's and CRIS's architecture is that CLIPSeg uses ViT [11] as the vision encoder while CRIS uses ResNet [15, 16]. And, while textual prompt tuning is used for both models, the visual and multimodal prompt tuning techniques are

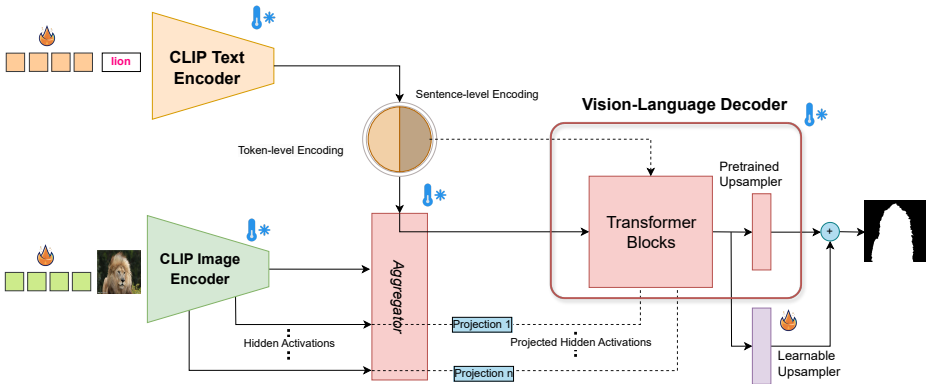


Fig. 3: Multimodal Prompt Tuning Architecture. To simplify, the projection layers for conditioning prompts from one mode to another are not shown here. Likewise, for unimodal techniques, only either of the prompt modalities is fed into the model.

applied only to CLIPSeg [32] because they require a ViT encoder. Thus, we have used six prompt tuning methods for CLIPSeg — three multimodal, one visual, and two textual, whereas two textual prompt tuning methods for CRIS. This amounts to a total of eight combinations of prompt tuning and VLSMs for each dataset.

Furthermore, inspired by Jia *et al.* [19], we have introduced a learnable upsampler (bottom right of Fig. 3) that takes the output of the penultimate layer of the decoder; its output is added to the decoder's output with learnable residual factor. This learnable block consists of a bilinear upsampler followed by a 2D convolution layer with kernel size 5×5.

6 Experiments

6.1 Datasets

For our empirical analysis of medical datasets, we utilize eight of the datasets and their splits provided by Poudel *et al.* [36]. These datasets include five non-radiology datasets — Kvasir-SEG [17], ClinicDB [5], and BKAI [3,35] for polyp segmentation in endoscopic images, DFU [22] for diabetic foot ulcer segmentation, and ISIC-16 [14] for skin lesion segmentation — and three radiology datasets — BUSI [2] for breast ultrasound segmentation, CAMUS [25] for 2D echocardiography segmentation, and CheXlocalize [40] for chest X-ray segmentation. For our analysis, we only use the foreground class name to learn the prompt for each dataset automatically.

As open-domain datasets, we choose Cityscapes [9] and PASCAL VOC 2012 [12] datasets consisting of 19 and 20 classes, respectively. The significance of using a natural-domain dataset is to evaluate the performance of various prompt-tuning methods in the same domain as the VLSMs were pretrained on. Also, we

Table 1: An overview of our datasets compared across the dimensions of category, type or modality, organ (for medical datasets), foreground classes, and their splits.

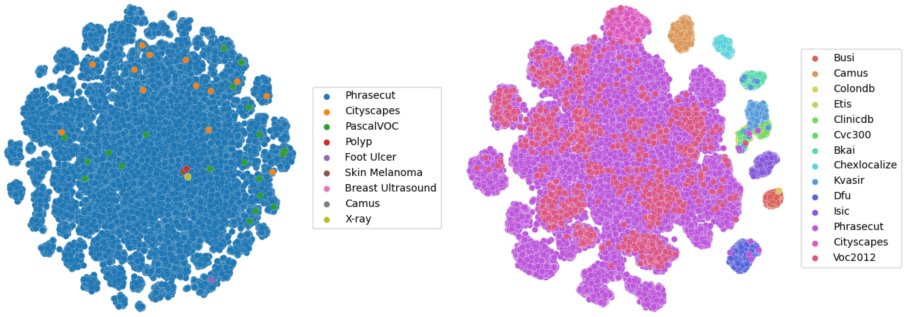
Category	Type	Organ	Name	Foreground Class(es)	# train/val/test
Non-Radiology	Endoscopy	Colon	Kvasir-SEG	Polyp	800/100/100
			ClinicDB		490/61/61
			BKAI		800/100/100
	Photography	Skin	ISIC 2016	Skin Lesion	810/90/379
		Foot	DFU 2022	Foot Ulcer	1600/200/200
Radiology	Ultrasound	Heart	CAMUS	Myocardium, Left ventricular, and Left atrium cavity	4800/600/600
		Breast	BUSI	Benign and Malignant Tumors	624/78/78
	X-Ray	Chest	CheXlocalize	Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumothorax, and Support Devices	1279/446/452
Open Domain	Street Scenes	N/A	Cityscapes	Bicycle, Building, Bus, Car, Fence, Motorcycle, Person, Pole, Rider, Road, Sidewalk, Sky, Terrain, Traffic Light, Traffic Sign, Train, Truck, Vegetation, Wall	34723/6005/-
	Diverse Scenes	N/A	PascalVOC	Aeroplane, Bicycle, Bird, Boat, Bottle, Bus, Car, Cat, Chair, Cow, Dining table, Dog, Horse, Motorbike, Person, Potted plant, Sheep, Sofa, Train, TV Monitor	2170/2148/-

can observe how the context learners perform when the number of classes (≈ 20) is higher compared to a maximum of 10 in the above-mentioned medical datasets. A detailed description of the datasets selected for this study is shown in Tab. 1.

To analyze the domain shift between open-domain and medical datasets, we have plotted t-SNE [33] of CLIP's text and vision embeddings for all datasets in Fig. 4. In addition to the datasets mentioned in Sec. 6.1, we have also plotted the phrases in the Phrasecut dataset [50], the one used to pretrain CLIPSeg. The figure has fewer text embeddings than image embeddings because many images correspond to a single prompt in the images in a dataset and the endoscopy datasets share the same foreground class (polyp). Since the number of phrases for the semantic segmentation datasets is insignificant in comparison to the referring image segmentation dataset like PhraseCut, we cannot draw many conclusions from Fig. 4a. However, in Fig. 4b, we can see the image embeddings for the medical datasets forming their clusters, separate from that of the open domain datasets. Similarly, the image embeddings of the open domain datasets overlap and the datasets with polyp as their foreground class are in the same cluster (center right). This illustrates a significant domain shift in CLIP's image embeddings from the dataset in which VLSMs were trained, to the medical datasets, thus, image embeddings require more adjustment to the domain shift.

6.2 Implementation Details

Before feeding the images to the models, we scaled and normalized the images in the original resolution of the respective models viz. 416×416 for CRIS [45] and 352×352 for CLIPSeg [32] using the bicubic interpolation. Additionally, for the training split, we slightly augmented the images by randomly scaling the images in the range of 2%, translating 2%, and rotating 5° using the Albumentations



(a) t-SNE of the Phrases **(b)** t-SNE of Images

Fig. 4: t-SNE [33] plots of phrases and images of all the datasets. Here, Phrasecut [50] is the dataset on which CLIPSeg was pretrained, Cityscapes [9] and Pascal VOC2012 [12] are the open-domain datasets, while others correspond to the medical domain. In Fig. 4b, we can see the overlap of clusters for the open domain datasets, and the medical datasets have formed separate small clusters.

library [7]. Additionally, the brightness and contrast of the images are augmented in the range of 10%.

We used 16-bit floating-point mixed-precision training in the NVIDIA GPUs using AdamW [31] optimizer with an effective batch size of 32 for all the experiments in this paper. For the loss function, we used a combined loss of Dice Loss [34] and Binary Cross Entropy loss. Mathematically, the loss function used can be shown as follows.

$$\mathcal{L} = \lambda_d \mathcal{L}_d + \lambda_{ce} \mathcal{L}_{ce}$$

Here, λ_d and λ_{ce} were chosen to be 1 and 0.2, respectively for all the datasets.

As Jia *et al.* [19] shown that the same set of hyperparameters like learning rate and weight decay may not be optimal for all the datasets, we performed

Table 2: The search space for the hyperparameter search.

Hyperparameter	**Search Space**	**Applicable for**	**Space Type**
Learning rate	$[10^{-5}, 5 \times 10^{-3}]$	ALL	Log
Weight decay	$[10^{-5}, 0.01]$	ALL	Log
Prompt depth	$[1, 11]$	ALL	Integer
Intermediate dimension	$\{32, 64, 96, 128\}$	CoCoOp, Maple	Choice
Use LORA	$\{true, false\}$	CoCoOp, Maple	Choice
Transformer: Number of Heads	$\{16, 20, 32\}$	Shared Attention	Choice
Transformer: Dropout Probability	$[0.1, 0.55]$	Shared Attention	Linear
Transformer: Feed-Forward Dim	$\{1280, 1420\}$	Shared Attention	Choice
Transformer: LayerNorm First	$\{true, false\}$	Shared Attention	Choice
Shared Space Dimension	$\{32, 64\}$	Shared Separate	Choice

hyperparameter sweeps using TPESampler [48] from Optuna [1] library. We ran each experiment 20 times with the search space for each parameter shown in Tab. 2. Although CoOp [62] and CoCoOp [61] have proposed to use prompt depth of 1, we wanted to observe how much can we benefit from increasing the prompt depth in those textual tuning techniques and to make the comparison across the prompting techniques fair and uniform, we have kept the prompt depth as a hyperparameter for all the context learners.

7 Results and Discussion

We report the results for each viable combination of models and context learns under the search space defined by Tab. 2 corresponding to the best set of hyperparameters in Tab. 3. We explain the results shown in the table in subsequent subsections.

7.1 Choice of Prompt Tuning Technique

From Tab. 3, we can see that the textual prompt tuning methods viz. CoCoOp and CoOp perform significantly worse than other prompt tuning techniques. The table shows that although Maple [23] has the highest dice scores for some datasets, VPT [19] has the highest average dice score across the test datasets despite not tuning the textual prompts. We hypothesize that since we used the same number of search trials for all the context learners, finding the optimal set of hyperparameters for VPT is easier as it has fewer hyperparameters than the multimodal variants (see Tab. 2). Also, this could have occurred because, as shown in Fig. 4, there is a significant domain shift from the open domain to the medical domain in the image domain only. This premise is supported by the observation that VPT performs better than the multimodal variants in the medical domain but the multimodal variants perform better than VPT in the

Table 3: The dice scores for each viable combination of VLSMs, and context learners, and end-to-end(E2E) finetuning for all datasets. Each cell corresponds to the best dice score on each dataset after searching 20 times using the search space defined by Tab. 2.

	Datasets →	Non-Radiology					Radiology			Open Domain		Overall
Model	Context Learner	BKAI	ClinicDB	Kvasir	DFU	ISIC	BUSI	CAMUS	Chexlocalize	Cityscapes	PascalVOC	Mean $\pm_{Std}$
CRIS	E2E finetune	92.40	91.69	91.39	76.13	91.94	69.31	91.09	62.57	63.50	75.73	-
CLIPSeg	E2E finetune	86.47	88.74	89.51	73.24	92.12	64.32	88.85	59.56	59.43	78.88	-
CLIPSeg	Maple	84.82	**90.61**	87.25	**72.68**	**92.10**	80.99	88.95	58.04	**56.70**	**79.48**	$79.16_{\pm 12.2}$
CLIPSeg	Shared Sep	86.06	90.30	88.02	71.71	92.03	81.71	89.29	**58.95**	**56.70**	79.46	$79.42_{\pm 12.22}$
CLIPSeg	Shared Attn	84.63	88.43	88.07	68.56	91.77	78.22	83.08	53.24	55.58	78.80	$77.04_{\pm 12.91}$
CLIPSeg	VPT	**87.96**	90.31	**89.03**	71.76	91.99	**82.45**	**89.36**	57.67	56.51	79.29	$\mathbf{79.63_{\pm 12.67}}$
CLIPSeg	CoCoOp	56.78	74.21	75.63	58.64	89.56	70.19	62.16	42.77	47.65	75.66	$65.33_{\pm 13.61}$
CLIPSeg	CoOp	53.85	66.59	73.53	54.64	89.00	67.28	58.92	41.82	47.92	75.50	$62.91_{\pm 13.49}$
CRIS	CoCoOp	76.85	85.09	82.12	58.24	86.87	72.95	81.96	51.28	44.88	72.46	$71.27_{\pm 14.03}$
CRIS	CoOp	75.55	76.73	80.38	57.18	86.10	75.46	79.23	51.79	46.50	71.72	$70.06_{\pm 12.69}$
											Average across Methods	$73.10_{\pm 6.24}$

Table 4: Performance of Maple when the first vector is initialized from Gaussian distribution vs 'a photo of a'.

Datasets →	Non-Radiology						Radiology		Open Domain		Overall
Init ↓	BKAI	ClinicDB	Kvasir	DFU	ISIC	BUSI	CAMUS	Chexlocalize	Cityscapes	PascalVOC	Mean ± Std
Gaussian	83.96	88.34	**88.65**	70.14	92.05	**81.25**	88.25	57.27	56.62	**79.50**	$78.6_{\pm 12.31}$
a photo of a	**84.82**	**90.61**	87.25	**72.68**	**92.10**	80.99	**88.95**	**58.04**	**56.70**	79.48	$\mathbf{79.16}_{\pm \mathbf{12.2}}$

open domain datasets (see Tab. 2). From this observation, we can infer that VPT is a good starting point when you have a dataset different from the ones on which the VLSMs were pretrained.

7.2 Importance of Prompt Initialization

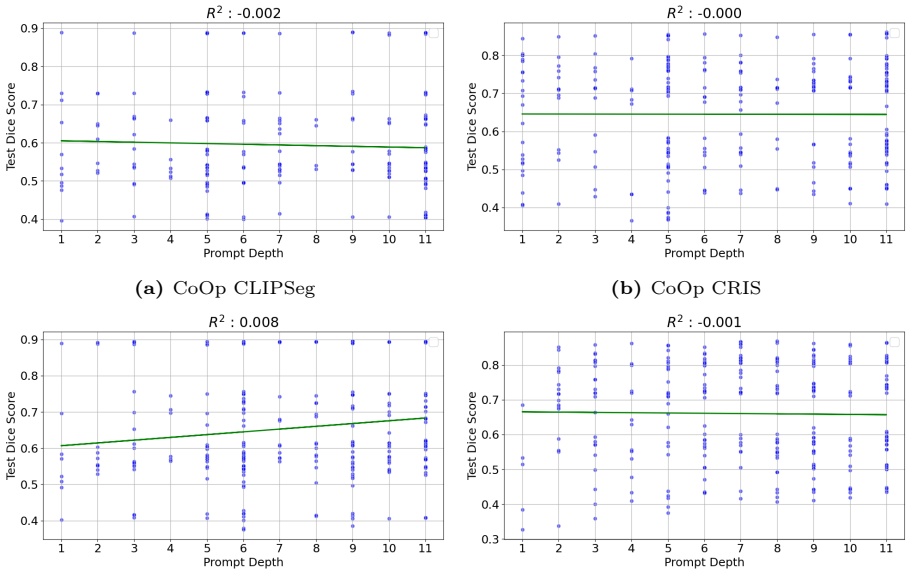
As shown by Zhou *et al.* [61, 62], a good initialization of the prompts leads to a better performance. Rather than randomly initializing the first prompt to CLIP's text encoder, they found initializing it with the embeddings of "a photo of a" led to a better performance. So, for the context learners where the prompts are initialized in the textual space viz. CoOp, CoCoOp, and Maple, the first prompt is initialized as mentioned above. We hypothesize that the ability to initialize the context vectors in the text space of the CLIP's text encoder in Maple [23] could be the reason why Maple performed better than the other multimodal prompt tuning techniques in some datasets. We trained Maple on CLIPSeg by initializing the context vectors from a Gaussian distribution ($\mu = 0, \sigma = 0.02$). From Tab. 4, we can see that the performance of Maple decreases for most of the dataset when the context vectors for the first depth are initialized at random, which supports our hypothesis. Additionally, the performance of different prompt tuning methods, is comparable for most of the datasets with end-to-end fine-tuning (Table 3, scores for the medical datasets are extracted from Poudel *et al.* [36] and the rest are trained with the same set of hyperparameters.).

7.3 Importance of Tuning Last Layer for Segmentation

As mentioned in Sec. 5, we have used a learnable upsampler as a residual connection to the decoder's last layer. To study the importance of the layer, we conduct an ablation by removing the learnable layer, keeping encoders and decoder frozen; the results from the ablation are shown in Tab. 5. From the table, we can see that the dice score has decreased for most of the cases, reducing the overall average dice score by 2.59. This shows the advantage of the learnable upsampler for segmentation.

7.4 Importance of Prompt Depth

Since prompt depth has shown to be an important hyperparameter of prompt tuning [19, 23, 53], we ran the experiments for various prompt depths as mentioned in Sec. 6.2. We have plotted the test dice score versus the prompt depth



(a) CoOp CLIPSeg **(b)** CoOp CRIS

(c) CoCoOp CLIPSeg **(d)** CoCoOp CRIS

Fig. 5: Test Dice *vs*. Prompt Depth for Textual Tuning of all Datasets

for various context learners for all the datasets in Fig. 5. The reason to include the dice score for all the datasets in the same graph is to observe if each prompt learning technique improves in general on increasing the prompt depth. From Figs. 5a to 5d, we can see that the textual context learners do not benefit significantly on increasing the prompt depth.

7.5 Learning Rate and Weight Decay

Similar to Jia *et al*. [19], we also observed that no single learning rate and weight decay pair performs the best for all the dataset and context learners (see

Table 5: The dice scores for each viable combination of VLSMs and context learners for all the datasets when a new learnable last upsampling layer is not introduced.

	Datasets →	Non-Radiology					Radiology			Open Domain		Overall
Models	Context Learner	BKAI	ClinicDB	Kvasir	DFU	ISIC	BUSI	CAMUS	Chexlocalize	Cityscapes	PascalVOC	Mean ± Std
CLIPSeg	Maple	81.56	85.55	86.45	66.38	91.20	80.95	84.96	55.96	53.16	76.95	$76.31_{\pm 12.58}$
	Shared Sep	80.45	83.31	86.07	65.46	91.00	78.81	85.71	57.72	53.06	76.72	$75.83_{\pm 12.14}$
	Shared Attn	80.21	85.89	85.62	62.23	90.91	76.27	84.30	55.77	51.86	76.50	$74.96_{\pm 12.92}$
	VPT	81.57	85.22	86.64	65.06	91.33	78.93	85.32	57.22	52.65	76.18	$76.01_{\pm 12.56}$
	CoCoOp	55.75	73.03	75.54	57.90	89.55	70.60	60.23	42.61	47.91	75.85	$64.9_{\pm 13.71}$
	CoOp	53.80	66.56	73.29	54.53	88.98	67.85	57.93	41.06	47.91	75.47	$62.74_{\pm 13.65}$
CRIS	CoCoOp	72.63	77.19	81.69	53.05	84.51	62.62	75.71	47.45	43.20	70.04	$66.81_{\pm 13.8}$
	CoOp	72.14	75.73	79.44	52.02	83.44	64.53	75.03	48.71	44.03	70.03	$66.51_{\pm 13.01}$
											Average across Methods	$70.51_{\pm 5.4}$

Appendix). So, we need to perform some sort of hyperparameter sweep to obtain the best set of configurations for the downstream dataset of choice.

8 Limitations

This work focuses on using VLSMs pretrained on segmentation tasks and evaluates their performance on downstream tasks. Since the classes used to train the models may not be the same as the ones used to evaluate the models in the downstream tasks, we chose VLSMs that output a binary mask, and the class to segment is provided through the text encoder. We refer to the VLSMs whose parameters do not need to be changed as the classes in downstream tasks differ from the pretrained task as *class-agnostic VLSMs*. Also, since the multimodal prompt tuning assumes the vision and text encoders run in parallel, one depending upon another, we restricted our study to VLSMs that use both encoders in (almost) parallel settings. Due to these constraints, we bounded our scope of study to CLIPSeg [32] and CRIS [45], VLSMs which output binary segmentation masks and use CLIP's encoders for the segmentation tasks.

9 Conclusion

In this paper, we propose an open-source benchmark framework, *TuneVLSeg*, to evaluate various unimodal and multimodal prompt tuning strategies on VLSMs for the segmentation task. Although VLSMs other than the ones used in this paper are also available, we used only the class-agnostic VLSMs because of the difference in the classes between the pretrained and the downstream datasets. Nevertheless, this can be extended to more VLSMs, if retaining the pretrained weights is not an issue, and more prompt tuning methodologies with minimum effort. Besides, to present the effectiveness of our framework, we report the performance of 6 prompt tuning methods on adapting 2 CLIP-based VLSMs to 2 natural domain and 8 diverse medical domain segmentation datasets. This work presents prompt tuning as an effective strategy to adapt pretrained foundational VLSMs for domain-specific segmentation tasks, with potential applications in clinical settings for medical image segmentation.

Acknowledgement. We would like to thank SUNY Korea for providing us with the computational resources required for this research project.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)

2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
3. An, N.S., Lan, P.N., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V.: BlazeNeo: Blazing fast polyp segmentation and neoplasm detection. IEEE Access **10**, 43669–43684 (2022)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
5. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
6. Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al.: An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247 (2024)
7. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020)
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
12. Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep **2007**(1-45), 5 (2012)
13. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19338–19347 (2023)
14. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin Lesion Analysis toward Melanoma Detection: A Challenge at ISBI 2016, hosted by ISIC. arXiv preprint arXiv:1605.01397 (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 558–567 (2019)
17. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-SEG: A segmented polyp dataset. In: MultiMedia Modeling. pp. 451–462. Springer (2020)

18. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
19. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
20. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423–438 (2020)
21. Jin, W., Cheng, Y., Shen, Y., Chen, W., Ren, X.: A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2763–2775 (2022)
22. Kendrick, C., Cassidy, B., Pappachan, J.M., O'Shea, C., Fernandez, C.J., Chacko, E., Jacob, K., Reeves, N.D., Yap, M.H.: Translating clinical delineation of diabetic foot ulcers into machine-interpretable segmentation. arXiv preprint arXiv:2204.11618 (2022)
23. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
24. Kwon, G., Cai, Z., Ravichandran, A., Bas, E., Bhotika, R., Soatto, S.: Masked vision and language modeling for multi-modal representation learning. In: The Eleventh International Conference on Learning Representations (2023)
25. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on medical imaging **38**(9), 2198–2210 (2019)
26. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059 (2021)
27. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597 (2021)
28. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: International Conference on Learning Representations (2021)
29. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
30. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68 (2022)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
32. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7086–7096 (2022)

33. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
34. 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). pp. 565–571. IEEE (2016)
35. Ngoc Lan, P., An, N.S., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V.: NeoUNet: Towards accurate colon polyp segmentation and neoplasm detection. In: Advances in Visual Computing. pp. 15–28. Springer (2021)
36. Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., Khanal, B.: Exploring transfer learning in medical image segmentation using vision-language models. In: Medical Imaging with Deep Learning (2023)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
38. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18082–18091 (2022)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., et al.: Benchmarking saliency methods for chest x-ray interpretation. Nature Machine Intelligence **4**(10), 867–878 (2022)
41. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4222–4235 (2020)
42. Shrestha, P., Amgain, S., Khanal, B., Linte, C.A., Bhattarai, B.: Medical vision language pretraining: A survey. arXiv preprint arXiv:2312.06224 (2023)
43. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022)
44. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
45. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
46. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision. pp. 631–648. Springer (2022)
47. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)

48. Watanabe, S.: Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv preprint arXiv:2304.11127 (2023)
49. Wu, C.E., Tian, Y., Yu, H., Wang, H., Morgado, P., Hu, Y.H., Yang, L.: Why is prompt tuning for vision-language models robust to noisy labels? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15488–15497 (2023)
50. Wu, C., Lin, Z., Cohen, S., Bui, T., Maji, S.: Phrasecut: Language-based image segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10216–10225 (2020)
51. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: CPT: Colorful prompt tuning for pre-trained vision-language models. AI Open **5**, 30–38 (2024)
52. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)
53. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225 (2022)
54. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
55. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022)
56. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and Medicine p. 108238 (2024)
57. Zhang, Y., Zhou, K., Liu, Z.: Neural prompt search. arXiv preprint arXiv:2206.04673 (2022)
58. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
59. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Li, X., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)
60. Zhong, Z., Friedman, D., Chen, D.: Factual probing is [mask]: Learning vs. learning to recall. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5017–5033 (2021)
61. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
62. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
63. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: ZegCLIP: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023)
64. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15659–15669 (2023)

65. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024)