

Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-supervised Medical Image Segmentation

Aishik Konwer^{1,2*}, Zhijian Yang^{1†}, Erhan Bas^{1†}, Cao Xiao¹, Prateek Prasanna³,
 Parminder Bhatia¹, Taha Kass-Hout¹

¹GE Healthcare

²Department of Computer Science, Stony Brook University

³Department of Biomedical Informatics, Stony Brook University

akonwer@cs.stonybrook.edu, prateek.prasanna@stonybrook.edu

{zhijian.yang, erhan.bas, cao.xiao, parminder.bhatia, taha.kass-hout}@gehealthcare.com

Abstract

Foundational models such as the Segment Anything Model (SAM) are gaining traction in medical imaging segmentation, supporting multiple downstream tasks. However, such models are supervised in nature, still relying on large annotated datasets or prompts supplied by experts. Conventional techniques such as active learning to alleviate such limitations are limited in scope and still necessitate continuous human involvement and complex domain knowledge for label refinement or establishing reward ground truth. To address these challenges, we propose an enhanced Segment Anything Model (SAM) framework that utilizes annotation-efficient prompts generated in a fully unsupervised fashion, while still capturing essential semantic, location, and shape information through contrastive language-image pretraining and visual question answering. We adopt the direct preference optimization technique to design an optimal policy that enables the model to generate high-fidelity segmentations with simple ratings or rankings provided by a virtual annotator simulating the human annotation process. State-of-the-art performance of our framework in tasks such as lung segmentation, breast tumor segmentation, and organ segmentation across various modalities, including X-ray, ultrasound, and abdominal CT, justifies its effectiveness in low-annotation data scenarios.

1. Introduction

With advancements in medical image analysis, there is an increasing need for sophisticated methods [20, 21, 29] to leverage the vast availability of radiology datasets (such as

X-ray, CT, and MRI) for accurate organ and tumor segmentation, as well as disease classification. The results of these tasks are crucial for physicians in designing effective treatment plans and surgical procedures. Several state-of-the-art deep learning-based foundational models, such as Vision-Language Models (VLMs), are now available for these purposes, relying on custom prompting to generate relevant predictions. However, they face two significant challenges: (1) despite being trained with only sparse prompts such as points or bounding boxes, these models still require human supervision for the prompt generation, leading to inefficiencies; and (2) many datasets lack comprehensive annotations, resulting in under-utilization during training of complex, data-hungry foundational architectures. Additionally, the high cost of human annotation efforts to create ground truth data can significantly escalate model development expenses.

Popular architectures such as SAM and CLIP, have been extended to medical data, and led to innovations such as BiomedCLIP, Merlin, SAM-Med2D, and SAM-Med3D [6, 11, 47, 55]. Recently, numerous studies have focused on enhancing SAM by replacing geometric prompts and integrating semantic and spatial knowledge in an unsupervised manner through techniques such as (1) self-prompting, (2) class activation maps from CLIP, and (3) object localization models [19, 24, 57]. As shown in Fig. 1, self-prompting (b) does not require expert-provided prompts during inference. On the other hand, unsupervised prompts (c) in [19] are only used for training SAM to generate pseudo-labels for weakly supervised downstream tasks. They lack sufficient location and shape information, which could be provided by text-based prompts. Hence, the question arises: *Can we come up with more refined and efficient prompts that can deliver stronger signals to foundational models without requiring intervention from annotators in both training and testing*

*Work done during internship at GE Healthcare

†Corresponding author

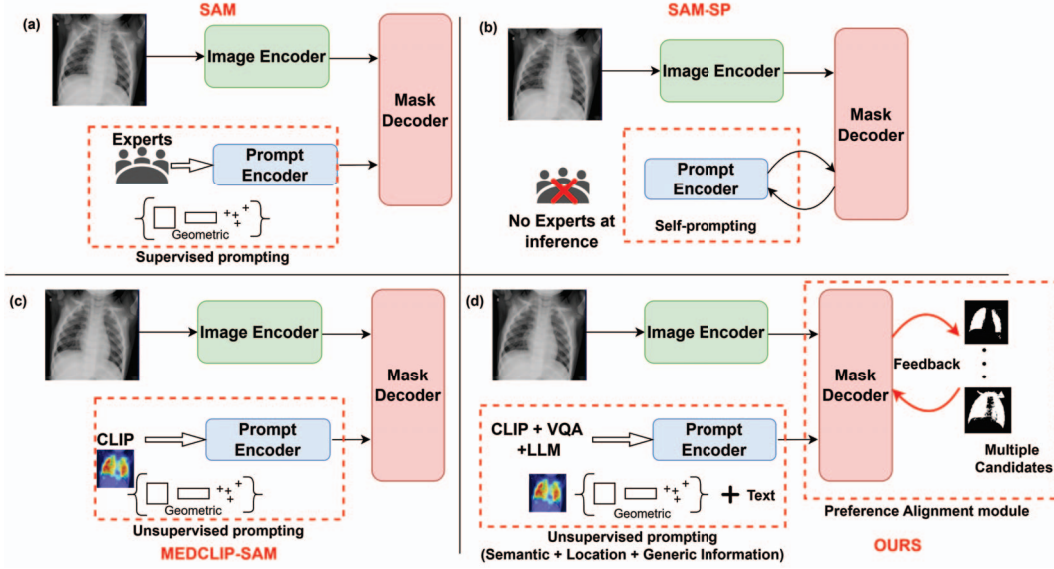


Figure 1. **Overview of our model:** (a) SAM and SAM-based approaches rely on expert prompts during both training and inference. (b) SAM-SP [57] introduces a self-prompting module, eliminating the need for expert prompts during inference. (c) MedCLIP-SAM [19] uses unsupervised semantic prompts to generate pseudo-labels through SAM. (d) Our approach not only combines semantic, location, and generic information via unsupervised prompts but also introduces a preference-based alignment module to reward or penalize the model.

stages? State-of-the-art models also lack comprehensive integration of semantic, locational, and generic class information in their prompting strategies. In our approach, we leverage CLIP, VQA, and LLM models [1, 55, 56] to extract this discriminative information, improving segmentation performance in unsupervised settings.

To tackle the challenges posed by limited annotated datasets, several weakly supervised semantic segmentation algorithms [2, 19] have been proposed. Single-stage methods [2] leverage coarse image labels to perform segmentation in an end-to-end fashion, while many two-stage approaches [19] utilize foundational models to generate pseudo labels that can be used to train downstream segmentation architectures. Additionally, human-in-the-loop frameworks [23] have proven beneficial for tasks such as image generation, segmentation, and prognosis, involving annotators to refine segmentation labels or assess the plausibility of synthetic images. This feedback can either be incorporated back into the model or used as ground truth to train a reward function via reinforcement learning (RL), significantly reducing the need for annotated training data. However, these methods still rely on explicit reward modeling, which prevents end-to-end training. In some cases, annotators must provide complex ratings or refinements, making it difficult to train a reward function efficiently. We ask another question: *Can we skip training a reward function and develop a straightforward, end-to-end pipeline that relies on simple annotator preferences?* To address this, we draw inspiration from direct preference optimization tech-

niques in the LLM preference alignment literature. We propose a preference-based reward model within our framework, using a novel loss function for end-to-end training. Our approach thus aligns with human preferences without the need for explicit reward modeling and is both easy to implement and train.

Our first goal is to develop refined and efficient prompts to fine-tune the SAM-Med2D encoder on diverse datasets, including 2D chest X-rays, breast ultrasound, and 3D abdominal CT scans. Initially, we input an image into the encoder alongside the BiomedCLIP [55] and MedVInT (VQA) [56] models. Corresponding texts, examples such as “Chest X-ray” and “Describe the condition of the lungs and location of pathologies,” are fed into BiomedCLIP and MedVInT, respectively. We gather generic information for the disease class from GPT-4, which is combined with the answers from the VQA model. Saliency maps generated by the CLIP model undergo dense CRF postprocessing to obtain bounding boxes. These bounding boxes and textual prompts are then inputted into the prompt encoder. Subsequently, the mask decoder receives both the image encoder embeddings and prompt embeddings to produce the segmentation maps.

After fine-tuning the prompting module on a small proportion of annotated data, we introduce our second major idea: simulating human feedback through a AI preference alignment module. We generate multiple segmentation candidates for a given image by thresholding the SAM output probabilities at various levels. These candidates are rated

on a scale of 1 to 4 based on the overlap between the candidates and ground truth, mimicking the evaluation process of a human annotator. This approach does not require explicit access to ground truth data for training, thus making our task a form of semi-supervised segmentation. We propose a DPO-inspired [38] loss function that encourages the model to prioritize desirable segmentation outputs by rewarding higher-rated candidates and penalizing lower-rated ones. The model is thus trained on the remaining portion of the dataset, without annotations, to perform relevant medical image segmentation tasks.

Overall our contributions can be summarized as follows:

- We propose refined and efficient unsupervised prompting strategies that deliver comprehensive information about location, semantics, and general disease/organ characteristics to our SAM-Med2D-based framework. Such an approach enhances segmentation performance while reducing reliance on human input for geometric prompts.
- We introduce a novel DPO-inspired loss function that facilitates semi-supervised model improvement using simulated human-in-the-loop feedback, eliminating the need for an explicit reward function. The framework generates multiple segmentation maps and rates them based on segmentation overlap, mimicking the evaluation of a human annotator. The model learns to distinguish between favorable and unfavorable candidates effectively.

2. Related Work

2.1. Vision-language models for medical domain

CLIP [37] has gained much popularity in medical image analysis. [14] fine-tuned CLIP on various PubMed articles to create PubMedCLIP. MedCLIP [48] leverages unpaired image and text datasets along with a semantic-matching loss to align visual and textual information. Windsor et al. [49] employ unimodal self-supervision, local-global contrastive losses, and data augmentation to enhance zero-shot performance and retrieval task efficiency in low-resource data settings. Some modality-specific CLIP variants [15, 54] have been developed for Chest X-ray and Mammograms due to the readily available image-text data in these areas. However, BiomedCLIP [55] stands out as the most recent model, excelling in scalability and performance across diverse multi-organ cross-modal retrieval tasks. Therefore, we have integrated BiomedCLIP into the CLIP-driven bounding box generation module of our framework.

Building on the success of large language models (LLMs) such as LLaMa [44] and GPT [35], researchers have explored merging visual features with textual representations using techniques such as cross-attention, Q-former, instruction tuning, and projection layers. This effort has resulted in vision-language models (VLMs) [4, 26, 30, 58], including Flamingo, BLIP-2, LLaVA, and

MiniGPT, which were further adapted for the medical domain through pretraining on multimodal medical datasets. Med-Flamingo [33] is the first medical visual question-answering (VQA) model with few-shot generation capabilities. RadFM [50] serves as a foundational model that also incorporates 3D volume data. Pretraining on extensive datasets, PMC-15M and PMC-VQA [55, 56], has facilitated the development of LLaVA-Med [25] and Med-VInT [56], respectively. Our segmentation pipeline leverages MedVInT’s capabilities for enhanced localization and shape-based answer generation related to tumors, organs, and disease manifestations.

2.2. SAM for multimodal biomedical data

MedSAM and SAM-Med2D [11, 31] focused on adapting SAM [18] for medical applications by fine-tuning it on 2D medical datasets, while SAM-Med3D [47] introduced alternative modules to accommodate 3D volumes. Efficient approaches, for example, AutoSAM [41] utilize trainable prompt encoders, whereas FastSAM3D [42] employs flash attention to accelerate inference. MedLSAM [24] proposed a localization framework to generate 3D bounding boxes as prompt input. However, most of these methods require ground truth data (bounding boxes or points) for training the prompt encoder, whereas we propose an unsupervised route for the same. We leverage ad-hoc VLM models such as CLIP, VQA, and GPT-4 together to propagate comprehensive information—encompassing semantics, location, and generic disease/organ information—that significantly enhances segmentation performance.

2.3. Human-in-the-loop feedback

Human-in-the-loop training paradigm in medical imaging has primarily focused on two areas: active learning [34], which identifies optimal data points for labeling to maximize performance, and interactive feedback [8] on model predictions to calibrate parameters. Examples include UI-Net, DeepIGeoS, and BIFseg [5, 45, 46], which utilize expert-provided scribbles or bounding boxes alongside geodesic transforms and graph-cut techniques to refine segmentation labels. Rao et al. [39] introduced IMIL, the first framework that assigns clinicians to guide data augmentations on mispredictions, emphasizing disease-relevant regions while eliminating irrelevant ones. Recently, the success of human feedback in instruction-tuning, and aligning large language model outputs through RL objective [52, 53], has led researchers to utilize human preferences to evaluate synthetic natural and histopathology images, thereby improving image-to-text models [23, 43] by training a reward function. Training a reward function requires the curation of dedicated human preference datasets and also prohibits the framework from operating in an end-to-end manner. Additionally, it often demands advanced do-

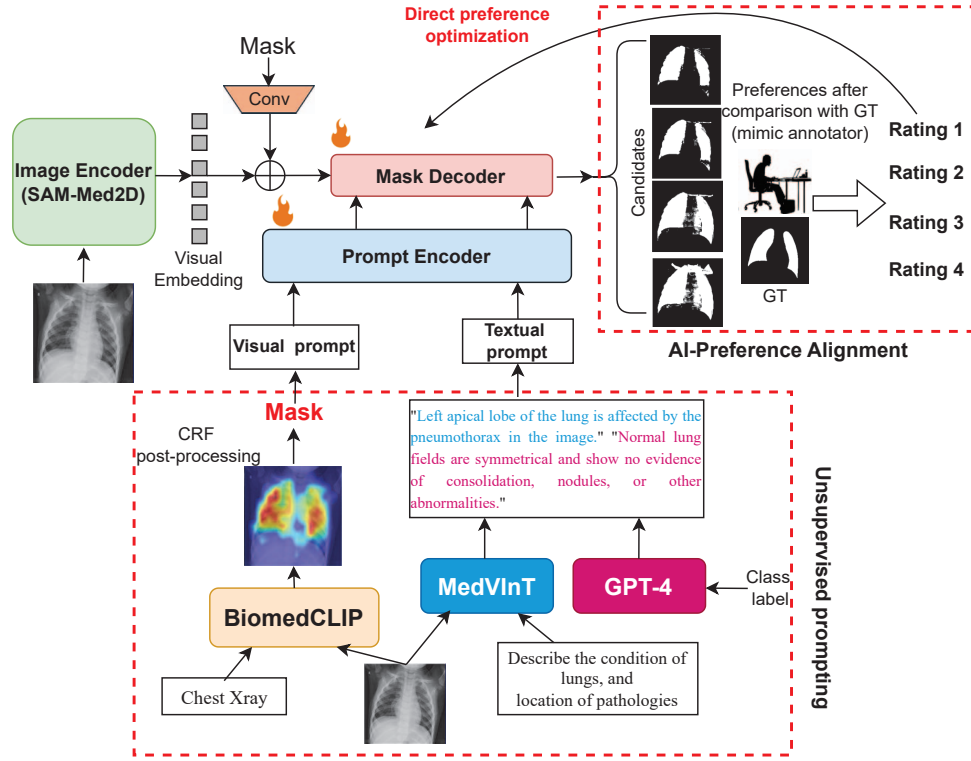


Figure 2. **Illustration of the proposed framework for semi-supervised segmentation:** Unsupervised geometric and text prompts, obtained from pretrained BiomedCLIP, MedVInT, and GPT-4 models, are fed into the prompt encoder for finetuning the framework on a small fraction of annotated data. In the next stage, we simulate a virtual annotation process that assigns ratings to the generated segmentation candidates, which are used to fine-tune the decoder. This stage handles unannotated data, as the model does not rely on ground truth for direct supervision but only for rating while simulating a human annotator’s feedback.

main knowledge from annotators, which can be costly. Our approach addresses this challenge by using direct preference optimization [38] to fulfill the RL objective, enabling the framework to serve as its own reward model. The model aligns generated proposals with appropriate preferences or ratings, leading to performance improvements, even with limited annotated data. These preferences are generated through a simulation mimicking human-in-the-loop feedback. Crucially, our method is easier to train and implement than traditional reward function-based RL pipelines.

3. Methodology

Overview. Given an input image, our aim is to generate prompts in the form of bounding boxes and textual queries, which will be used to produce segmentation masks. The image is processed by three pretrained components: the SAM-Med2D encoder, BiomedCLIP, and MedVInT. Concurrently, GPT-4 curates generic information about the concerned disease or organ. This information, combined with the BiomedCLIP-generated box prompt and the MedVInT-

provided textual answer prompt, serves as input for the prompt encoder. The prompt generation process is detailed in Sec. 3.1. Next, both the encoded image and prompt embeddings are fed into the decoder to create segmentation maps. Following this initial fine-tuning on annotated data, we introduce a novel loss to mimic human-in-the-loop feedback. By thresholding the output probabilities, our model produces four different segmentation candidates, and we incorporate insights about the quality of these candidates into the framework through the proposed loss function. This part, trained with unannotated data, is elaborated in Sec. 3.2.

3.1. Visual and textual prompt generation

Visual prompt. To generate visual prompts, we first input the image along with corresponding text prompts (e.g. “chest x-ray”, “benign breast tumor”, “left kidney”, “liver”, etc.) into BiomedCLIP, a foundational model pretrained on millions of medical image-caption pairs. Next, we leverage gScoreCAM [10] to create a saliency map highlighting targeted regions (organs or tumors) in the image corre-

sponding to the supplied text. These saliency maps are then post-processed with a conditional random field filter [22] to produce coarse segmentation masks. We apply an area constraint, retaining the largest connected component or up to two, depending on the dataset (the lung dataset may yield two components). Finally, we identify the bounding box coordinates within these closed components for the box prompt, while randomly sampling several points from the designated area for the point prompt.

Textual prompt. We extract visual embeddings from a given image by processing it through a vision encoder derived from the PMC-CLIP architecture [28]. This encoder features a pretrained ResNet50 [16] and a trainable projection layer constructed with stacks of transformer decoder blocks. Next, we create a prompt template that incorporates the question for the image, formatted as “Question: {}, Answer is:”. This prompt is then passed through a tokenization embedding layer initialized with the weights of PMC-LLaMA [51] to generate the text embedding. Finally, we concatenate the visual and text embeddings to form the input space for a pre-initialized multimodal transformer decoder. The answer generated from this VQA setup delivers essential information regarding the shape and location of anatomical structures and pathologies. Sample questions included in the prompt are: “What is the shape of the liver and where is it located?”, and “What is the shape of breast tumor and where is it located?” Additionally, we prompt GPT-4 with the relevant organ or disease label to obtain a generic description of its characteristics. Finally, both types of textual prompts are concatenated and provided as input to the prompt encoder, described in the following paragraph.

Prompt encoder and mask decoder. The prompt encoder, same as the one in SAM, supports three types of prompts: point, box, and text. Point and box prompts are represented by their positional encodings—specific coordinates for points, and the top-left and bottom-right corners for boxes—along with learnable feature representations. Text prompts are processed through a pretrained BiomedCLIP encoder to generate corresponding text embeddings. All prompt embeddings are then projected into 256-dimensional vectors. The feature map from the first iteration of the model, is downsampled through multiple convolutional layers followed by GELU activation to match the 256-channel dimension. Finally, these downsampled masks are combined element-wise with visual encoder embeddings. The mask decoder takes both the prompt embeddings and the visual embeddings to produce a segmentation map. The architecture is illustrated in Fig. 2.

3.2. Segmentation proposal generation and Preference Alignment

Proposal generation. After fine-tuning our framework, comprising the prompt encoder and mask decoder, using

ground truth for a fraction (10%) of the dataset, we propose integrating AI-based preferences for the next training episode. First, we generate multiple segmentation proposals by sampling different thresholds from the pixel probability scores output by the model. We then simulate an expert annotator to evaluate the quality of these proposals, assigning ratings based on the overlap between each generated proposal and the corresponding ground truth. Although ground truth is not explicitly used in this training phase, it plays a passive role by enabling the simulation of preferences in the absence of an actual human annotator in our experiments. We also experiment with two alternative rating mechanisms: one inspired by SAM, which backpropagates loss solely for the best candidate, and another that ranks all proposals rather than merely rating them. After obtaining the ratings or rankings, we fine-tune our decoder using the direct preference optimization technique to better align the segmentation outputs with the preferences of the virtual annotator.

RLHF Preliminaries. Language models typically utilize a reward function to align their generated responses with user preferences. This preference distribution is often modeled using the Bradley-Terry [7] model, which operates on pairwise comparisons. For a given prompt X , when the language model produces two responses, one more favorable, Y_m , and the other less favorable, Y_l , the distribution can be expressed as:

$$P(Y_m > Y_l | X) = \frac{e^{r^*(X, Y_m)}}{e^{r^*(X, Y_m)} + e^{r^*(X, Y_l)}}, \quad (1)$$

Parameters of the reward model can be estimated via Maximum Likelihood Estimation (MLE) and the goal is to minimize the below loss function:

$$\mathcal{L}_{\mathcal{R}} = -\mathbb{E}_{(X, Y_m, Y_l) \sim \mathcal{D}} [\log \sigma(r_{\phi}(X, Y_m) - r_{\phi}(X, Y_l))], \quad (2)$$

where σ is the logistic function and \mathcal{D} denotes the dataset of preferences.

Reinforcement Learning from Human Feedback (RLHF) is a widely used method that involves training a reward model based on user ratings. The primary goal is to discover an optimal policy with parameter π_{θ} that maximizes this reward function while incorporating a KL divergence term to ensure the model’s outputs do not deviate significantly from the original policy π_{ref} . The optimization problem can be formulated as follows:

$$\begin{aligned} & \max_{\pi_{\theta}} \mathbb{E}_{X \sim \mathcal{D}, Y \sim \pi_{\theta}(Y|X)} [r_{\phi}(X, Y)] \\ & - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta}(Y | X) || \pi_{ref}(Y | X)], \end{aligned} \quad (3)$$

Direct Preference optimization. Due to the discrete nature of language generation, this objective is not differentiable, which usually necessitates optimization through

reinforcement learning. The primary language model then leverages this framework to align its outputs with user ratings, ultimately generating high-scoring responses. In contrast, Direct Preference Optimization (DPO) is a more recent and streamlined approach that eliminates the need for a separate reward model. Instead, it fine-tunes the main language model directly using user preferences. This is achievable because DPO focuses on optimizing the policy itself rather than the reward function. The maximum likelihood objective for a parameterized policy π_θ in such a case can be formulated as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(X, Y_m, Y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(Y_m | X)}{\pi_{\text{ref}}(Y_m | X)} - \beta \log \frac{\pi_\theta(Y_l | X)}{\pi_{\text{ref}}(Y_l | X)} \right) \right], \quad (4)$$

where β is the weight applied to reward or penalize the responses. In DPO, models are fine-tuned using pairs of outputs, explicitly comparing a preferred response Y_m with less preferred ones Y_l . We adapt this concept for our application in the imaging domain, where multiple segmentation candidates Y_1, \dots, Y_4 are generated for the same {image, prompt} pair I . With ratings available for each candidate, we assign weights β_1, β_2 to these candidates (from best to worst) to reward the higher-quality segmentations Y_1, Y_2 and penalize the less desirable ones Y_3, Y_4 . Consequently, Equation 4 is modified to create Equation 5. This approach allows us to incorporate real-world annotator preferences into our method through a straightforward yet effective loss function. The parameters of the initially fine-tuned model from Sec. 3.1 are denoted as π_{fine} . Our objective is to find the optimal parameters π_ψ for the final architecture without deviating much from π_{fine} .

$$\mathcal{L}_{\text{DPO}}(\pi_\psi; \pi_{\text{fine}}) = -\mathbb{E}_{(I, Y_1, Y_2, Y_3, Y_4) \sim \mathcal{D}} \left[\log \sigma \left(\beta_1 \log \frac{\pi_\psi(Y_1 | I)}{\pi_{\text{fine}}(Y_1 | I)} + \beta_2 \log \frac{\pi_\psi(Y_2 | I)}{\pi_{\text{fine}}(Y_2 | I)} - \beta_2 \log \frac{\pi_\psi(Y_3 | I)}{\pi_{\text{fine}}(Y_3 | I)} - \beta_1 \log \frac{\pi_\psi(Y_4 | I)}{\pi_{\text{fine}}(Y_4 | I)} \right) \right] \quad (5)$$

4. Experimental Design and Results

Datasets. We evaluated our framework for semi-supervised segmentation using three public datasets covering lung, breast tumor, and abdominal organ segmentation across X-ray, ultrasound, and CT. For Ultrasound Breast Tumor Segmentation, we used 810 images from BUSI [3] and UDIAT [9], with 600 for training and 210 for testing. For Chest X-ray Lung Segmentation, we used 27,132 images from COVID-QU-Ex [12], spanning normal, lung opacity, viral pneumonia, and COVID-19 cases. For Abdominal CT Organ Segmentation, 200 annotated CT scans from the AMOS-CT training set were used for training. Full dataset details, including train-test splits, are in the supplementary Sec. 6.

Implementation Details. Our method is implemented in PyTorch [36] on an EC2 instance (with 64 GB NVIDIA T4 Tensor Core GPUs). For feature extraction, we utilize the SAM-Med2D pretrained encoder. First we use annotations for only 10% of the training dataset, during which all components (visual encoder, prompt encoder, and mask decoder) are fine-tuned. In this step, only the unsupervised prompting strategy is employed. Bounding box and point prompts are used together for all experiments. The remaining portion of the dataset was used in an unannotated form to train the DPO-driven alignment strategy. We optimize the model using the Adam optimizer [13], training for 15 epochs for prompt module fine-tuning and 30 epochs for alignment. In both stages, the initial learning rate is set to 1e-4 and is halved every 10 epochs. All images are resized to a resolution of 256×256 using the same resizing strategy as SAM-Med2D. Similar to our framework, we have also fine-tuned SOTA SAM-based methods on medical imaging datasets using LoRA by adding low-rank matrices to transformer layers. This enables efficient adaptation with fewer parameters, enhancing domain-specific performance while preserving pretrained features. The loss function used during the initial fine-tuning is a 20:1 weighted combination of focal loss [27] and Dice loss [32]. Details on the preference alignment module and proposal generation are provided in the supplementary material Sec. 6. Dice Similarity Coefficient, Intersection over Union (IoU), and Surface Dice Similarity (SDC) scores are used to evaluate segmentation performance.

4.1. Comparison with state-of-the-art

Quantitative results. Fig. 3 compares the performance of our framework with relevant methods (U-Net [40], nnU-Net [17], SAM [18], SAM-Med2D [11], Self-prompt). The self-prompt method is designed as a variant of [57], excluding the knowledge distillation module. We also design a prompt-only baseline of our framework which is trained on different splits of fully annotated data. Our framework is initially trained with only the prompting module using 10% of data (annotated). As a result, both the prompt-only and final versions exhibit identical performance on this 10% subset. For the final model, the additional data used for training are considered unannotated, as training the alignment mechanism does not use supervision from ground truth. SOTA methods directly use ground truth for the entire data subset. Nevertheless, our method consistently outperforms them in limited data settings (10-50%) due to the preference alignment mechanism. Our architecture thus reduces the reliance on large datasets, highlighting its lower annotation requirements, which makes it highly effective in low-data regimes. On the Chest-Xray dataset, for instance, with just 20% of the data, our method achieves a Dice score of 78.87, compared to 58.66 (U-Net), 60.97 (nnU-Net), 61.64 (SAM),

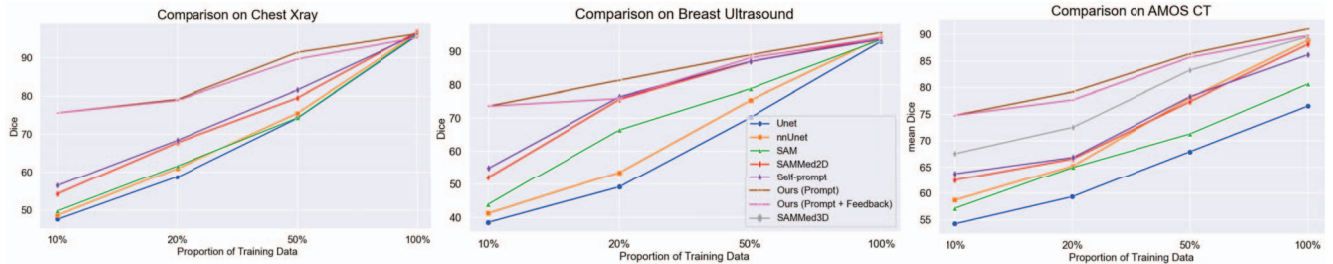


Figure 3. Quantitative comparison with SOTA. Dice score (for Chest Xray, Breast USD) and mean Dice score (for AMOS CT) have been shown to measure the model segmentation performance on different proportions of training data (10%, 20%, 50%, and 100%).

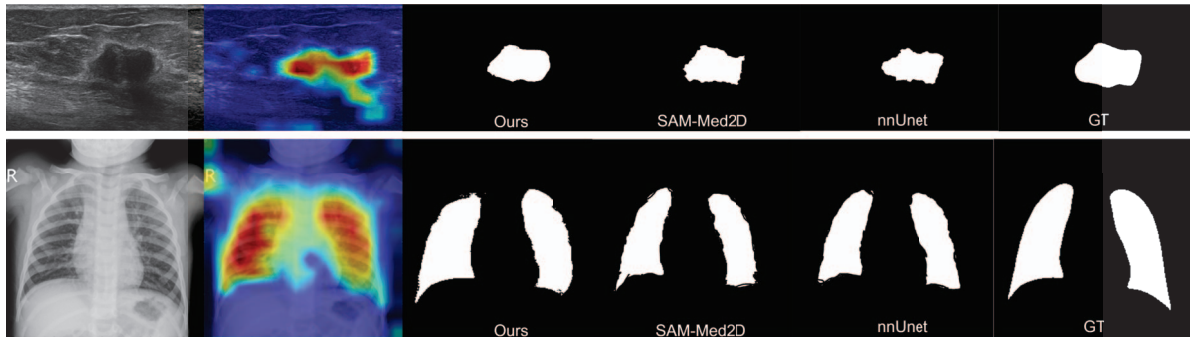


Figure 4. Qualitative comparisons were made between the segmentation results of nnUnet, SAM-Med2D, and our framework on 2D datasets. BiomedCLIP-based saliency maps are also depicted. Experiments were conducted in 50% data settings.

67.81 (SAM-Med2D), and 68.41 (Self-prompt). Both U-Net and nnU-Net are known to be data-hungry models, struggling with smaller data subsets. Our usage of visual and textual prompts, derived through BiomedCLIP and VLM models, offers superior signals to those used in SAM and SAM-Med2D. Unlike these SAM variants, Self-Prompt SAM does not require expert-supplied prompts during inference; it instead generates prompts from the output masks in each iteration. This method offers a slight improvement (+1%) over SAM-Med2D. Our method shows more stable performance gains across different data subsets, while other SOTA methods exhibit steeper improvements. In the 50% data setting, our semi-supervised framework achieves a Dice score of 89.68, compared to 91.42 for the supervised prompt-only version. Similar result trends are observed in the breast ultrasound dataset. Notably, the performance jump of our method from 20% to 50% data is much more pronounced than from 10% to 20%. This can be due to the nature of the dataset, as the model requires more data to achieve precise and accurate segmentation of breast tumors. We also evaluated our method on a 3D abdominal organ segmentation dataset, including organs such as the liver, kidneys, spleen, pancreas, aorta, bladder, etc (see Fig. 3). Our method outperforms the SOTA across all data proportions except for the full dataset. With 20% of the data, our method achieves a mean Dice score of 77.69, surpassing U-

Net (59.35), nnU-Net (65.21), SAM (64.93), SAM-Med2D (66.57), SAMMed3D (72.54), and Self-prompt (71.83). At the 50% data setting, it reaches a Dice score of 85.70, comparable to the 86.36 achieved by the prompt-only version.

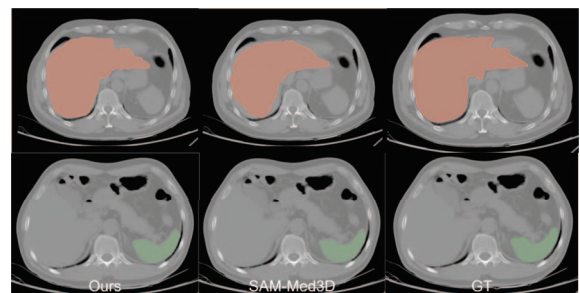


Figure 5. Segmentation maps of different anatomical structures (liver and spleen) for SAM-Med3D and our method

Qualitative results. We present segmentation maps for 3 datasets in Fig. 4 and Fig. 5, generated by our model trained at 50% data settings. The saliency maps could correctly highlight the target regions and proved to be an excellent source of supervision. It can be noted from Fig. 4 that our segmentation quality around the boundaries of tumor or lung is much superior compared to both nnUnet and SAM-Med2D. In Fig. 5, SAM-Med3D tended to under-segment the edges of both the abdominal organs, the liver,

Supervision	Methods	Chest-Xray		Breast-USD		AMOS-CT	
		IoU	Dice	IoU	Dice	mDice	mSDC
10% + 10% unannotated	Ours	74.30	78.87	64.35	75.88	77.69	78.34
20%	- Alignment	75.02	79.13	67.51	81.38	79.20	80.66
10%	- Alignment	68.43	75.60	61.44	73.62	74.77	76.06
	- Alignment - VQA	66.90	73.35	57.60	70.53	74.08	75.41
	- Alignment - VQA - GPT4	63.16	72.76	54.26	68.89	73.16	74.89
	- Alignment - VQA - CAM	50.38	57.02	45.16	59.05	69.97	71.45

Table 1. Ablation results demonstrating effectiveness of major components.

Supervision	Preference Alignment	Chest-Xray		Breast-USD		AMOS-CT	
		IoU	Dice	IoU	Dice	mDice	mSDC
10% + 10% unannotated	Loss for best candidate	70.37	77.09	61.76	73.81	75.01	76.36
	Rating	73.99	78.41	63.97	75.52	77.23	77.98
	Ranking	74.30	78.87	64.35	75.88	77.69	78.34
10% + 20% unannotated	Ranking	79.74	85.15	73.56	84.23	80.54	81.95
10% + 40% unannotated	Ranking	88.96	89.68	85.92	88.15	84.30	85.70

Table 2. Ablation results on different proportions of training data for 3 types of preference scoring strategies.

and spleen. Additional results are in supplementary Sec. 1.

4.2. Ablation studies

Effectiveness of major components. We conduct several ablation experiments (see Tab. 1) to evaluate the contribution of each module in our architecture. For simplicity, we focus on the Chest-Xray dataset for analysis. Before incorporating the DPO-driven alignment strategy, all baselines were trained with 10% labeled data. First, we developed a naive baseline (last row), which relies solely on textual answers generated from GPT-4. As expected, it performed poorly, achieving a 57.02 Dice score. A second baseline (second-last row) was designed to use only visual prompts from BiomedCLIP, which provides semantic information. This method significantly benefited from the coarse segmentation masks derived from CLIP saliency maps, improving the Dice score to 72.76. Next, we combined both textual and visual prompts to form a third baseline (-Alignment-VQA), which resulted in a slight improvement (+0.59%) over the CLIP-only baseline. Finally, we integrated VQA into the prompting strategy to obtain answers related to the shape and location of the target regions, completing our fully empowered unsupervised prompting strategy (-Alignment). This model achieved a Dice score of 75.60 with 10% data and improved to 79.13 with 20% data. In comparison, with the preference alignment mechanism, fine-tuning the prompting module with 10% annotated data and training the alignment module on an additional 10% unannotated data achieved a Dice score of 78.87. This is on par with the fully supervised prompt-only model, underscoring the effectiveness of our alignment module.

Preference Scoring strategies. We conducted several ablation studies (detailed in Tab. 2) to evaluate the effective-

ness of the preference-scoring strategy. One baseline approach involved backpropagating the loss based only on the best candidate, while another compared ranking the candidates rather than simply rating them. The results are presented for different proportions of unannotated data, on top of the foundational fine-tuning of both the prompt encoder and the decoder using 10% annotated data. At 10% unannotated data, both the ranking and rating approaches performed similarly (mean Dice scores of 77.69 and 77.23 for AMOS CT, respectively), significantly outperforming the best-candidate-only method (75.01). Additionally, we observed that as the proportion of unannotated data increased, the model’s performance improved substantially.

Robustness to noisy rating. We experimented with introducing noise into the rating mechanism by flipping the ratings of closely ranked candidates to enhance the framework’s robustness. Results are in supplementary material Sec. 2.

5. Conclusion

We introduce a novel training strategy to enhance SAM representations for semi-supervised medical image segmentation. We extract integrated semantic, location, and shape information from pretrained vision-language models in an unsupervised manner. This information is used as refined prompts for our model. We also implement an optimal policy, inspired by direct preference optimization in language models. This enables human-in-the-loop feedback simulation within a streamlined training framework, without the need for a separate reward function or extensive knowledge from annotators. These modules ensure that our framework outperforms state-of-the-art methods across datasets spanning multiple modalities in low-annotation data scenarios.

6. Acknowledgments

We sincerely thank everyone from the GE Healthcare Science & Technology Organization for their support and contributions to the project. The primary model development and evaluations were conducted during a summer internship at GEHC. After the internship, Konwer and Prasanna were supported by the following grants during the paper writing and rebuttal process: NIH 1R01CA297843-01, NIH 3R21CA258493-02S1, and 1R21CA258493-01A1.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [2] Peri Akiva and Kristin Dana. Towards single stage weakly supervised semantic segmentation. *arXiv preprint arXiv:2106.10309*, 2021. 2
- [3] Walid Al-Dhabyani, Mohammed Goma, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 6, 2
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3
- [5] Mario Amrehn, Sven Gaube, Mathias Unberath, Frank Schebesch, Tim Horz, Maddalena Strumia, Stefan Steidl, Markus Kowarschik, and Andreas Maier. Ui-net: Interactive artificial neural networks for iterative image segmentation based on a user model. *arXiv preprint arXiv:1709.03450*, 2017. 3
- [6] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truys, et al. Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512*, 2024. 1
- [7] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 5
- [8] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71: 102062, 2021. 3
- [9] Michal Byra, Piotr Jarosik, Aleksandra Szubert, Michael Galperin, Haydee Ojeda-Fournier, Linda Olson, Mary O’Boyle, Christopher Comstock, and Michael Andre. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomedical Signal Processing and Control*, 61:102027, 2020. 6, 2
- [10] Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. gscorecam: What objects is clip looking at? In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 1959–1975, 2022. 4
- [11] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyang Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 1, 3, 6
- [12] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 6, 2
- [13] P Kingma Diederik. Adam: A method for stochastic optimization. (*No Title*), 2014. 6
- [14] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023. 3
- [15] Shantanu Ghosh, Clare B Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 632–642. Springer, 2024. 3
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [17] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 6
- [18] 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*, pages 4015–4026, 2023. 3, 6
- [19] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 643–653. Springer, 2024. 1, 2
- [20] Aishik Konwer, Xuan Xu, Joseph Bae, Chao Chen, and Prateek Prasanna. Temporal context matters: Enhancing single image prediction with disease progression representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18824–18835, 2022. 1
- [21] Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21415–21425, 2023. 1
- [22] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *Inter-*

- national conference on machine learning*, pages 513–521. PMLR, 2013. 5
- [23] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 2, 3
- [24] Wenhui Lei, Xu Wei, Xiaofan Zhang, Kang Li, and Shaoting Zhang. Medlsam: Localize and segment anything model for 3d medical images. *arXiv preprint arXiv:2306.14752*, 2023. 1, 3
- [25] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3
- [27] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 6
- [28] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023. 5
- [29] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [31] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3
- [32] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6
- [33] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3
- [34] Vishwesh Nath, Dong Yang, Bennett A. Landman, Daguang Xu, and Holger R. Roth. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10): 2534–2547, 2021. 3
- [35] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3
- [36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [38] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4
- [39] Adrit Rao, Andrea Fisher, Ken Chang, John Christopher Panagides, Katherine McNamara, Joon-Young Lee, and Oliver Aalami. Imil: Interactive medical image learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5241–5250, 2024. 3
- [40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 6
- [41] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023. 3
- [42] Yiqing Shen, Jingxing Li, Xinyuan Shao, Blanca Inigo Romillo, Ankush Jindal, David Dreizin, and Mathias Unberath. Fastsam3d: An efficient segment anything model for 3d volumetric medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 542–552. Springer, 2024. 3
- [43] Shenghuan Sun, Greg Goldgof, Atul Butte, and Ahmed M Alaa. Aligning synthetic medical images with clinical knowledge using human feedback. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [45] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7): 1562–1573, 2018. 3

- [46] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018. 3
- [47] H Wang et al. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. *Preprint at <https://arxiv.org/abs/2310.15161>*, 2024. 1, 3
- [48] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 3
- [49] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Vision-language modelling for radiological imaging and reports in the low data regime. *arXiv preprint arXiv:2303.17644*, 2023. 3
- [50] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 3
- [51] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024. 5
- [52] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 3
- [53] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023. 3
- [54] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 3
- [55] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 2, 3
- [56] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 2, 3
- [57] Chunpeng Zhou, Kangjie Ning, Qianqian Shen, Sheng Zhou, Zhi Yu, and Haishuai Wang. Sam-sp: Self-prompting makes sam great again. *arXiv preprint arXiv:2408.12364*, 2024. 1, 2, 6
- [58] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3