

MedPixel: A Unified Pixel-Language Model for Medical Reasoning and Segmentation

Haoyu Yang¹, Meixing Shi¹, Zengjie Chen¹, Haoran Sun²
Haitao Leng³, Xiaoming Shi⁴, Yuxiang Cai¹, Yankai Jiang⁵

¹Zhejiang University ²Fudan University ³Kuaishou
⁴East China Normal University ⁵Shanghai Artificial Intelligence Laboratory
{yanghaoyu, shimeixing, caiyuxiang}@zju.edu.cn; jyk1996ver@zju.edu.cn

Abstract

Reliable medical image understanding requires models to connect clinical language and visual reasoning with pixel-level grounding. Yet medical vision-language models often lack precise localization, whereas medical segmenters typically rely on explicit target categories or precise spatial prompts. This divide is reinforced by a supervision mismatch: segmentation datasets provide precise masks but little language supervision, whereas medical vision-language data rarely pair language with dense spatial annotations. To address this gap, we present **MedPixel**, a unified medical pixel-language model built around a shared language-mask interface. To provide scalable supervision, we introduce **MedPLG-440K**, comprising approximately 440K pixel-language task samples constructed through a clinically motivated synthesis process without external LLM annotation. MedPixel is trained with joint multi-task supervised fine-tuning followed by **Pixel-Level Preference Optimization**, which uses ground-truth masks as offline verifiers to derive response preferences from mask quality. MedPixel supports a broad spectrum of tasks spanning explicit grounding, implicit reasoning, spatial interaction, grounded explanation, and medical VQA. Across this task spectrum, MedPixel achieves strong performance in both pixel-level prediction and response generation, together with effective zero-shot transfer to external grounding benchmarks and robustness to imperfect spatial prompts. Code and model checkpoints will be released at <https://github.com/yhy-whu/Medpixel>.

1 Introduction

Medical image analysis is evolving from isolated recognition and segmentation tasks toward general systems that can support broader clinical decision making. Clinicians synthesize imaging findings with anatomical knowledge, clinical context, and diagnostic intent to identify relevant abnormalities, interpret their significance, and determine where they occur. This evolution expands the role of medical models from processing fixed image inputs to understanding images, language, and spatial prompts, and from producing fixed outputs to supporting reasoning, interaction, localization, and explanation. We summarize these practical requirements into five representative medical task formats: explicit grounding, implicit reasoning, spatial interaction, grounded explanation, and medical visual question answering. Together, these interactions reflect a broader trend toward medical systems

that integrate clinical understanding with spatially grounded visual analysis.

Current medical models emphasize different aspects of this functionality. Mainstream medical vision-language models focus primarily on image-level understanding and text generation, including medical question answering, report generation, and clinical reasoning (Li et al. 2023; Sellergren et al. 2025; Bannur et al. 2024). Medical segmentation models provide accurate pixel-level predictions across diverse imaging modalities, but usually rely on predefined target categories or explicit spatial prompts (Kirillov et al. 2023; Ravi et al. 2025; Ma et al. 2024, 2025; Cheng et al. 2023; Zhao et al. 2024).

Recent pixel-language models, including LISA, PixelLM, SAM4MLLM, and UniPixel, connect multimodal language understanding with dense prediction through shared language-mask representations, enabling natural-language grounding, reasoning segmentation, and visual-prompt interaction (Lai et al. 2024; Ren et al. 2024; Chen et al. 2024c; Liu et al. 2026). Related medical models extend pixel-level grounding to biomedical images and increasingly support complex or implicit clinical queries that require target inference before localization (Huang et al. 2025a; Wu et al. 2025; Huang et al. 2025b; Yan et al. 2026). Despite this progress, unified support for diverse medical pixel-language interactions remains limited.

As the capability space broadens, supervision becomes a new bottleneck. Different forms of language-pixel interaction often require different types of supervision. This challenge is particularly pronounced in medicine, where existing data are highly asymmetric. Segmentation datasets provide abundant pixel-accurate masks but little language supervision, whereas medical vision-language datasets contain rich questions and answers but rarely pair them with dense spatial annotations. Although recent efforts have expanded biomedical datasets with spatial grounding and reasoning supervision (Wu et al. 2025; Deng et al. 2025; Huang et al. 2025b; Yan et al. 2026), constructing dedicated supervision for each interaction form remains costly. The central challenge is therefore not simply to add more capabilities, but to obtain diverse medical pixel-language supervision in a scalable manner.

This motivates us to reconsider the role of existing segmentation masks. A mask is conventionally treated as the target

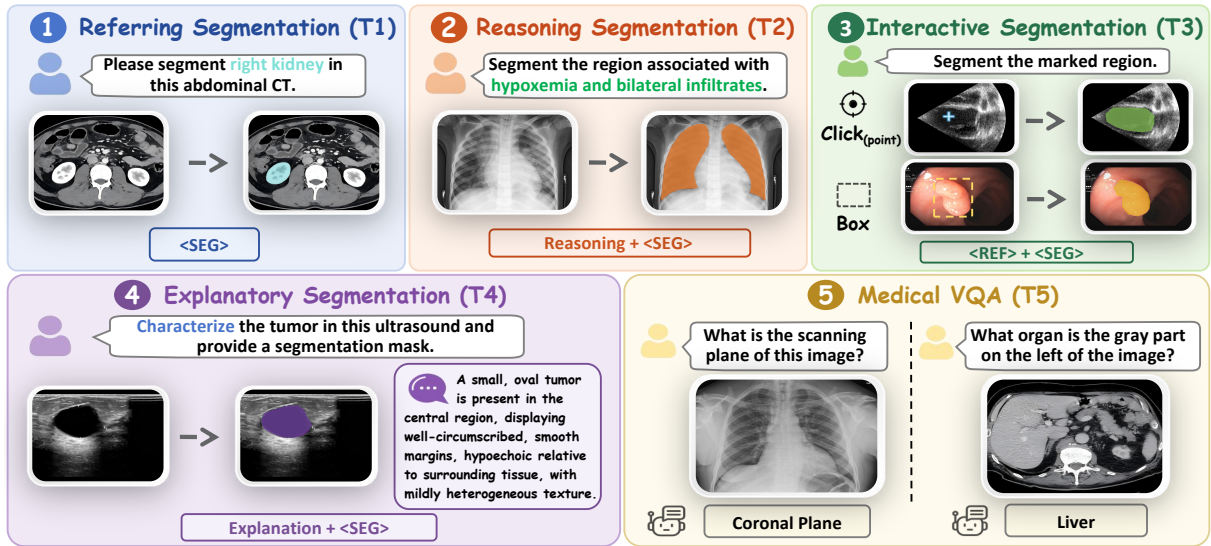


Figure 1: Representative task interfaces supported by MedPixel. T1–T4 produce pixel-grounded outputs through the `<SEG>` interface, with T3 additionally incorporating point or box guidance through `<REF>`; T5 produces text-only answers. MedPixel unifies referring, reasoning, interactive, and explanatory segmentation with medical VQA in a single framework.

of a segmentation objective, yet the corresponding annotated region also provides a spatial anchor for connecting image content with language. Together with image, category, and modality information, it can support descriptions of where a finding appears, how it is visually characterized, and what target is being referred to. Inspired by clinical image analysis, these cues can be organized from visual observation and characterization toward target interpretation and spatial grounding. Existing medical masks therefore constitute an underused source of structured pixel-language supervision.

Based on this insight, we construct **MedPLG-440K** (**Medical Pixel-Language Grounding 440K**), comprising approximately 440K pixel-language task samples across four grounded interaction formats. These samples are synthesized from existing medical segmentation annotations through a clinically motivated process without external LLM annotation. We separately incorporate medical VQA data to preserve general image-level understanding. Figure 1 illustrates the resulting task spectrum.

Built on this supervision, we introduce **MedPixel**, a unified medical pixel-language model based on Qwen2.5-VL and SAM2 (Bai et al. 2025; Ravi et al. 2025). A shared language–mask interface connects language generation with dense mask decoding through a special `<SEG>` token, enabling diverse interaction forms to share a common backbone and segmentation pathway.

The shared language–mask interface also creates an alignment problem. The hidden representation of each generated `<SEG>` token directly conditions mask decoding, yet standard supervised fine-tuning optimizes reference responses and target masks without comparing the spatial consequences of alternative generations. Responses that are similarly plausible in language may therefore produce different `<SEG>` representations and masks of substantially different quality,

creating a mismatch between response likelihood and pixel-level correctness. We introduce **Pixel-Level Preference Optimization (PLPO)** to address this mismatch. Ground-truth masks act as task-native offline verifiers that rank candidate responses by the quality of their induced masks. These rankings are converted into response preferences, aligning language generation with pixel-level outcomes without requiring an external reward model.

Another practical consideration is the quality of user-provided spatial prompts. Promptable segmentation methods are typically trained and evaluated with precise point or box prompts, whereas real users may provide only approximate spatial guidance. As prompt quality decreases, segmentation performance can degrade substantially (Kirillov et al. 2023; Ravi et al. 2025; Ma et al. 2024, 2025; Huang et al. 2024). By combining semantic descriptions with spatial cues, MedPixel can use language to clarify the intended target when the spatial prompt is incomplete or inaccurate.

Our contributions are summarized as follows:

- We introduce **MedPLG-440K**, comprising approximately 440K pixel-language samples across four grounded interaction formats, synthesized without external LLM annotation.
- We present **MedPixel**, a unified model for four pixel-level interactions and medical VQA, together with **PLPO** for aligning language generation with mask quality.
- MedPixel achieves leading localization and response-generation performance, with effective zero-shot transfer and robustness to imperfect box prompts.

2 Related Work

2.1 Medical Vision-Language and Segmentation Models

Medical vision-language models have advanced medical image understanding through multimodal pretraining and instruction tuning. Representative models such as LLaVA-Med, MedGemma, and MAIRA-2 support medical question answering, report generation, and clinical reasoning (Li et al. 2023; Sellergren et al. 2025; Bannur et al. 2024). In parallel, foundation segmentation models such as SAM and SAM2 have inspired medical adaptations including MedSAM, MedSAM2, and SAM-Med2D, while Biomed-Parce extends category-driven segmentation across diverse biomedical modalities (Kirillov et al. 2023; Ravi et al. 2025; Ma et al. 2024, 2025; Cheng et al. 2023; Zhao et al. 2024). These two directions provide complementary strengths: vision-language models offer flexible semantic interaction, whereas segmentation models provide precise spatial outputs. This complementary development has motivated models that connect language understanding with pixel-level grounding.

2.2 Pixel-Level Grounding and Medical Reasoning

Pixel-level vision-language models connect language understanding and dense prediction. LISA introduces a special segmentation token that maps language-model representations to masks, enabling reasoning segmentation from natural-language instructions (Lai et al. 2024). PixelLM extends this paradigm to multi-object grounding, SAM4MLLM incorporates visual prompts for richer multimodal interaction, and UniPixel unifies referring, mask generation, visual prompting, and object-centric reasoning (Ren et al. 2024; Chen et al. 2024c; Liu et al. 2026). Together, these methods make pixel-level prediction an integral part of multimodal interaction and reasoning rather than merely a final output.

This direction has also been extended to medical imaging. MedPLIB and UniBiomed combine medical language understanding with pixel-level prediction across diverse tasks and imaging settings (Huang et al. 2025a; Wu et al. 2025). MedSeg-R reasons over complex clinical instructions beyond explicit target descriptions, while MedReasoner further addresses implicit queries that require target inference before localization (Huang et al. 2025b; Yan et al. 2026). Building on this progression, MedPixel unifies multiple grounded medical interaction formats, while MedPLG-440K systematically derives their supervision from existing segmentation annotations rather than constructing separate supervision for each format.

3 Method

Figure 2 summarizes the overall design of MedPixel. The framework consists of three parts: (a) MedPLG-440K construction, (b) unified MedPixel architecture, (c) two-stage training strategy. We first build MedPLG-440K from existing medical segmentation annotations, then use it to support a shared language-mask interface for unified text generation

and dense prediction, and finally optimize the model with joint multi-task supervised fine-tuning followed by PLPO.

3.1 Task Interfaces and Architecture

Task Formulation. Given a medical image I , a task instruction q , and, for interactive segmentation, an optional spatial reference r , MedPixel generates a textual response y and, when required, a segmentation mask $\hat{M}$. As summarized in Figure 1, the first four tasks produce mask-grounded responses, whereas medical VQA produces text-only answers. **Language-Mask Interface.** MedPixel couples Qwen2.5-VL (Bai et al. 2025) with a SAM2 (Ravi et al. 2025) mask decoding branch. Qwen2.5-VL performs multimodal understanding and language generation, while the mask branch produces dense segmentation outputs. All interaction formats are processed through the same multimodal backbone. Segmentation-oriented responses contain a special $\langle \text{SEG} \rangle$ token, while interactive inputs additionally contain a spatial reference token $\langle \text{REF} \rangle$.

For segmentation-oriented tasks, MedPixel generates a special $\langle \text{SEG} \rangle$ token at the position where mask prediction is required. Let $h_{\langle \text{SEG} \rangle}$ denote the hidden state of this token. A learnable projection Φ_{seg} maps it into the SAM2 prompt embedding space:

$$z_{\text{seg}} = \Phi_{\text{seg}}(h_{\langle \text{SEG} \rangle}). \quad (1)$$

Given the SAM2 image embedding F_I , the mask decoder predicts

$$\hat{M} = \mathcal{D}_{\text{SAM2}}(F_I, z_{\text{seg}}), \quad (2)$$

where $\mathcal{D}_{\text{SAM2}}$ denotes the mask decoding pathway. This shared language-mask interface allows different query and response formats to use a common dense prediction mechanism.

Spatial Reference Injection. Interactive segmentation incorporates point or box guidance through a special $\langle \text{REF} \rangle$ token. Given a spatial reference r , a trainable prompt encoder converts it into a sparse spatial representation, which is then projected to the language-model hidden dimension:

$$z_{\text{ref}} = \Phi_{\text{ref}}(\mathcal{E}_{\text{ref}}(r)), \quad (3)$$

where $\mathcal{E}_{\text{ref}}$ denotes the spatial prompt encoder and Φ_{ref} the projection into the Qwen2.5-VL embedding space. Before the sequence is processed by the language model, z_{ref} replaces the original token embedding at the $\langle \text{REF} \rangle$ position. Spatial guidance therefore enters the same multimodal sequence as image and text inputs, allowing the subsequent $\langle \text{SEG} \rangle$ representation to integrate semantic and spatial information before mask decoding.

3.2 MedPLG-440K Construction

As illustrated in Figure 2(a), MedPLG-440K is constructed from existing medical segmentation annotations. Each source sample provides a medical image I , a ground-truth mask M , a target category c , and metadata m . Based on these annotations, we first perform visual attribute extraction to obtain a visual descriptor d , and then synthesize four segmentation-oriented interaction formats:

$$(I, M, c, m) \xrightarrow{\mathcal{A}} d \xrightarrow{\mathcal{T}} \{(x_t, y_t, M)\}_{t=1}^4, \quad (4)$$

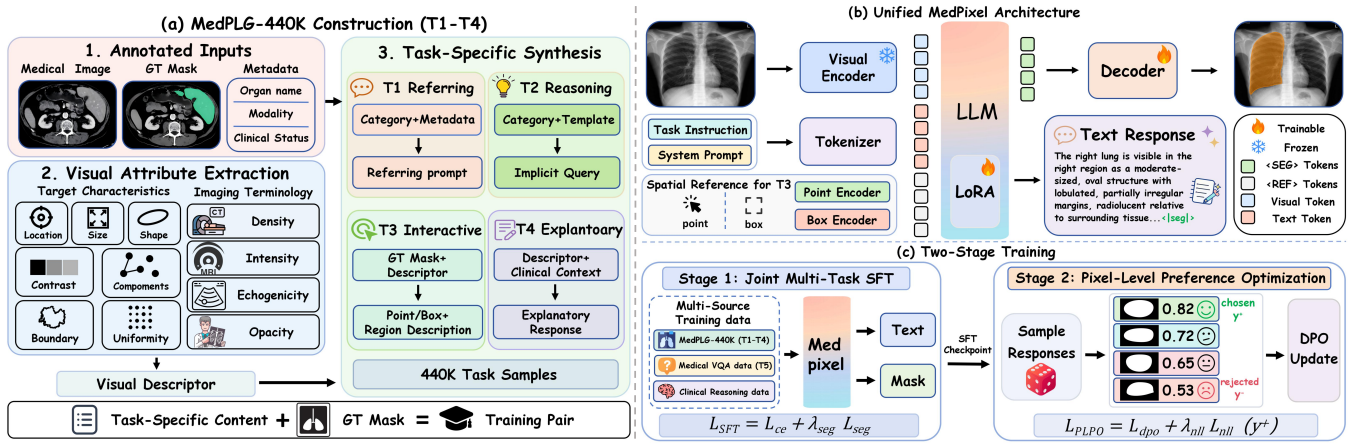


Figure 2: Overview of MedPixel. (a) Construction of MedPLG-440K. (b) MedPixel architecture for unified language generation and mask prediction. (c) Two-stage training with multi-task SFT and PLPO.

where $\mathcal{A}$ denotes visual attribute extraction and $\mathcal{T}$ denotes task-specific synthesis.

Visual Attribute Extraction. For each annotated region, we derive a structured visual descriptor from mask geometry and image statistics. The descriptor captures location, size, shape, contrast, component structure, boundary, and uniformity. Location and size are obtained from the mask centroid and relative area; shape and component structure are characterized by region geometry and connected components; boundary regularity is estimated from contour measurements. Contrast and uniformity are computed from differences between the foreground and background and from variation within the region. Appearance attributes are expressed using terminology appropriate to the imaging modality. For example, appearance characteristics may be described in terms of density for CT, signal intensity for MRI, echogenicity for ultrasound, or opacity for X-ray. This process yields a compact visual descriptor aligned with the annotated region.

Task-Specific Synthesis. The source annotations and visual descriptor are combined according to the target interaction format. For *T1 referring segmentation*, the target category and metadata are used to construct explicit referring prompts. For *T2 reasoning segmentation*, category information is paired with curated templates to form implicit queries that require target inference before localization. For *T3 interactive segmentation*, points or bounding boxes are sampled from the ground-truth mask, while the visual descriptor is converted into a concise region description. For *T4 explanatory segmentation*, the visual descriptor is combined with concise clinical context to construct an explanatory response grounded in the target region. Each synthesized sample is paired with the original ground-truth mask to form a pixel-language training pair.

Through this pipeline, we construct **MedPLG-440K**, comprising approximately 440K pixel-language task samples. During Stage 1, MedPLG-440K is further combined with medical VQA and auxiliary clinical reasoning data to support joint training across five tasks. Dataset sources and construction details are provided in Appendices A and B.

3.3 Two-Stage Training

Figure 2(c) summarizes the two-stage optimization procedure of MedPixel. Stage 1 learns the shared language-mask interface through joint multi-task supervised fine-tuning. Stage 2 further aligns response generation with mask quality through PLPO.

Stage 1: Joint Multi-Task SFT. Stage 1 jointly trains MedPixel on a unified mixture of all tasks. The objective combines autoregressive language modeling with mask supervision:

$$\mathcal{L}_{SFT} = \mathcal{L}_{ce} + \lambda_{seg} \mathcal{L}_{seg}, \quad (5)$$

where $\mathcal{L}_{ce}$ denotes the language modeling loss and $\mathcal{L}_{seg}$ denotes the pixel-level segmentation loss. For samples with segmentation supervision, the mask objective is

$$\mathcal{L}_{seg} = \lambda_{focal} \mathcal{L}_{focal} + \lambda_{dice} \mathcal{L}_{dice} + \lambda_{iou} \mathcal{L}_{iou} + \lambda_{obj} \mathcal{L}_{obj}. \quad (6)$$

For samples without mask supervision, the mask term is omitted. This joint objective allows the five task formats to share the same language-mask interface while learning both language generation and pixel-level prediction.

Stage 2: Pixel-Level Preference Optimization.

Building on the Stage 1 checkpoint, we construct response preference pairs from reasoning segmentation samples according to the quality of the masks induced by alternative generations. For each input x_i , we sample N candidate responses from π_{SFT} and decode the mask associated with each generated $\langle \text{SEG} \rangle$ representation. The quality of the n -th candidate is measured against the corresponding ground-truth mask:

$$s_i^{(n)} = \text{Dice} \left(\hat{M}_i^{(n)}, M_i^{\text{gt}} \right), \quad (7)$$

where $\hat{M}_i^{(n)}$ denotes the mask induced by the n -th candidate response. Ground-truth masks are used only for offline scoring and are never exposed during candidate generation. This process converts pixel-level outcomes into response preference supervision without requiring an additional reward model or online rollouts.

We discard empty responses and candidates that do not produce a valid $\langle \text{SEG} \rangle$ output from which a mask can be decoded. Among the remaining candidates, the highest-scoring response is selected as the chosen response y_i^+ , while a lower-scoring valid response is selected as the rejected response y_i^- . To avoid weak or ambiguous preference supervision, we retain a pair only when

$$s_i^+ \geq \tau_{\min}, \quad s_i^+ - s_i^- \geq \tau_{\text{gap}}, \quad (8)$$

where $\tau_{\min}$ ensures sufficient quality of the chosen response and τ_{gap} enforces a meaningful quality difference between the chosen and rejected responses. Inputs that do not satisfy these criteria are discarded. The resulting preference dataset is

$$\mathcal{P} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^{|\mathcal{P}|}. \quad (9)$$

Given $\mathcal{P}$, we optimize MedPixel using Direct Preference Optimization (Rafailov et al. 2023). Let π_θ denote the trainable policy and $\pi_{\text{ref}} = \pi_{\text{SFT}}$ the frozen reference policy. We define the relative preference log-ratio as

$$\Delta_\theta = \log \frac{\pi_\theta(y^+ | x)}{\pi_{\text{ref}}(y^+ | x)} - \log \frac{\pi_\theta(y^- | x)}{\pi_{\text{ref}}(y^- | x)}. \quad (10)$$

The preference objective is

$$\mathcal{L}_{\text{PLPO}} = -\mathbb{E}_{\mathcal{P}} [\log \sigma(\beta \Delta_\theta)] + \lambda_{\text{nll}} \mathcal{L}_{\text{nll}}(y^+ | x), \quad (11)$$

where β controls the preference strength and the NLL term anchors the model to the chosen response.

During Stage 2, the segmentation projector Φ_{seg} and the SAM2 mask decoder remain frozen. Preference optimization therefore reshapes the response distribution and the resulting $\langle \text{SEG} \rangle$ representations while preserving the dense decoding pathway learned in Stage 1. In this way, MedPixel converts differences in pixel-level mask quality into response-level preferences, encouraging generations whose $\langle \text{SEG} \rangle$ representations induce more accurate segmentation outcomes.

4 Experiments

4.1 Experimental Setup

Training data. Stage 1 trains MedPixel on a 1.84M-sample mixture of medical VQA (64%), segmentation (32%), and clinical reasoning (4%) data. MedPLG-440K provides the four segmentation-oriented tasks (T1–T4), which are incorporated through task-specific sampling and replication. The VQA component is aggregated from public medical QA datasets (Chen et al. 2024b; Rückert et al. 2024; Zhang et al. 2023; Liu et al. 2021; Jin et al. 2019; Pal, Umapathi, and Sankarasubbu 2022; Jin et al. 2021; Chen et al. 2025), while the reasoning component combines HuatuoGPT-o1, Medical-R1-Distill, Citrus-S3 (Chen et al. 2024a; Freedom-Intelligence 2025; Wang et al. 2025a), and image-grounded reasoning synthesized from PubMedVision (Chen et al. 2024b). The resulting checkpoint is used to mine offline preference pairs for Stage 2. Full data statistics and training settings are provided in Appendix C.

Evaluation Sets. T1–T4 are derived from the held-out official BiomedParse test split, which is excluded from training. T1 evaluates explicit referring segmentation on 24,391 representative slices selected to reduce redundancy in volumetric

studies. T2 and T4 contain 2,500 samples each for implicit reasoning and grounded explanation, respectively. T3 contains 2,000 interactive samples, with point and box prompts derived from ground-truth masks. T5 contains 10,267 image-based and text-based multiple-choice questions from public medical QA benchmarks. We further evaluate external transfer on the test sets of MeCoVQA-G+ and U-MRG-14K without benchmark-specific training.

Metrics. For T1–T4, we report Dice and normalized surface Dice (NSD) with a 5-pixel tolerance; both are expressed as percentages, with the “%” symbol omitted. For T2 and T4, we additionally report METEOR (Banerjee and Lavie 2005) for response similarity. For T2, we also report target grounding accuracy (TGA), which measures whether the inferred target concept is fully recovered. Empty or invalid responses receive zero scores. For T5, we report multiple-choice accuracy. MeCoVQA-G+ is evaluated using slice-level micro and modality-level macro Dice, NSD, and HD95, while U-MRG-14K is evaluated using Dice. Full evaluation protocols are provided in Appendix D.

Baselines. We compare MedPixel with pixel-language models, segmentation specialists, and medical VQA models. Pixel-language baselines include LISA (Lai et al. 2024) and LISA++ (Yang et al. 2023), PixelLM (Ren et al. 2024), SAM4MLLM (Chen et al. 2024c), VisionReasoner (Liu et al. 2025), UniBiomed (Wu et al. 2025), Citrus-V (Wang et al. 2025b), MMedAgent (Li et al. 2024), MedPLIB (Huang et al. 2025a), and UniPixel (Liu et al. 2026). Segmentation specialists include BiomedParse (Zhao et al. 2024), MedSAM (Ma et al. 2024), MedSAM2 (Ma et al. 2025), SAM2 (Ravi et al. 2025), and SAM-Med2D (Cheng et al. 2023). For T5, we additionally compare with LLaVA-Med (Li et al. 2023), Qwen2.5-VL (Bai et al. 2025) and HuatuoGPT-Vision (Chen et al. 2024b).

Implementation details. We train MedPixel at 3B and 7B scales using Qwen2.5-VL as the multimodal backbone. LoRA (Hu et al. 2022) is applied to the $q/k/v/o$ attention projections of both the language model and Qwen vision tower, using $r=128$ and $\alpha=256$, while their base weights remain frozen. Stage 1 optimizes the language adapters, multimodal projector, special token embeddings, segmentation projector, spatial reference modules, and SAM2 branch. Stage 2 updates the response-generation modules while freezing the segmentation branch.

4.2 Main Results

Table 1 compares MedPixel with segmentation specialists, pixel-language models, and medical VQA baselines. MedPixel achieves strong performance throughout the evaluated interfaces. On T1 referring segmentation, MedPixel-7B reaches 85.0 Dice and 61.7 NSD, exceeding BiomedParse by 5.0 and 6.7 points, respectively, showing that the unified framework retains strong explicit grounding performance. The advantage becomes substantially larger when localization requires semantic target inference, with improvements of 29.2 Dice on T2 reasoning segmentation and 40.7 Dice on T4 explanatory segmentation over the strongest baselines. MedPixel also achieves the highest METEOR on both tasks and the strongest TGA on T2, showing that the localiza-

Method	Size	T1 Refer.		T2 Reason.				T3 Interact.		T4 Explan.			T5 VQA	
		Dice	NSD	Dice	NSD	MET.	TGA	Dice	NSD	Dice	NSD	MET.	IMCQ	TMCQ
Segmentation specialists														
MedSAM	93M	–	–	–	–	–	–	43.6	32.9	–	–	–	–	–
MedSAM2	38M	–	–	–	–	–	–	75.7	52.4	–	–	–	–	–
SAM2	0.2B	–	–	–	–	–	–	61.7	31.9	–	–	–	–	–
SAM-Med2D	0.3B	–	–	–	–	–	–	58.6	30.2	–	–	–	–	–
BiomedParse	0.3B	80.0	55.0	–	–	–	–	–	–	–	–	–	–	–
Pixel-language and VQA models														
UniBiomed	1B	65.0	27.7	28.3	14.4	4.9	27.3	–	–	35.8	19.1	7.3	7.0	6.6
Citrus-V	8B	45.7	27.9	37.5	25.9	–	–	–	–	35.1	23.0	10.2	41.1	51.1
VisionReasoner	7B	42.4	14.6	33.4	14.5	29.7	39.8	–	–	31.1	13.0	17.2	41.6	47.0
LISA++	7B	34.7	7.8	22.8	8.9	7.9	0.1	–	–	12.1	5.3	12.0	–	–
SAM4MLLM	8B	27.1	3.5	20.8	3.6	–	–	–	–	21.4	4.2	–	–	–
LISA	7B	26.9	7.7	26.3	13.2	–	–	–	–	23.0	9.0	–	–	–
MMedAgent	7B	27.8	7.9	25.6	12.8	–	–	–	–	25.4	12.5	–	6.7	8.8
MedPLIB	7B	17.1	4.4	8.3	3.1	–	–	–	–	11.1	3.8	–	36.4	31.0
PixelLM	7B	30.6	3.3	18.2	3.0	8.1	4.9	–	–	10.0	1.5	11.6	–	–
UniPixel	7B	39.0	9.6	27.9	7.5	3.0	0.0	26.6	7.2	30.2	9.5	3.2	–	–
LLaVA-Med	7B	–	–	–	–	–	–	–	–	–	–	–	28.5	37.1
Qwen2.5-VL	3B	–	–	–	–	–	–	–	–	–	–	–	38.1	43.4
HuatuoGPT-Vision	7B	–	–	–	–	–	–	–	–	–	–	–	42.0	46.7
MedPixel	3B	84.3	60.4	62.7	49.5	50.2	64.1	76.0	52.1	73.9	52.2	40.4	42.9	52.5
MedPixel	7B	85.0	61.7	66.7	53.4	51.3	70.8	75.9	53.1	76.5	55.2	42.3	47.5	58.1

Table 1: Comparison with representative methods across five tasks. MET. denotes METEOR, and TGA denotes target grounding accuracy. IMCQ and TMCQ denote image- and text-based multiple-choice accuracy, respectively. “–” denotes unsupported native interfaces. For T3, results are averaged over point and box protocols.

tion gains are achieved while preserving response similarity and target identification. On T3, MedPixel reaches 76.0 Dice under its native interface combining semantic and spatial cues, remaining competitive with specialized promptable segmenters while supporting a broader interaction format. On T5, MedPixel-7B surpasses the strongest baselines by 5.5 and 7.0 points on image- and text-based MCQs, respectively.

Scaling from 3B to 7B further improves MedPixel’s overall capability, with gains across most tasks and particularly clear improvements on T2, T4, and T5. T1 improves modestly, while T3 remains essentially unchanged. This pattern suggests that increased model capacity strengthens semantic interpretation, target inference, and response generation, while the spatially guided interactive pathway is already strong at the 3B scale. Scaling enhances higher-level medical understanding without compromising dense grounding ability. Representative examples are shown in Figure 1.

4.3 External Generalization

MeCoVQA-G+. We evaluate zero-shot transfer on MeCoVQA-G+, which contains 2,719 grounding samples across eight medical modalities. The benchmark is not used during training, and all results in Table 2 are reproduced under the same evaluation protocol. Since MedPLIB is trained on MeCoVQA-G+, we report it only as a benchmark-trained reference rather than a zero-shot baseline. MedPixel-7B achieves the strongest overall performance among zero-shot text-driven models, improving over BiomedParse by +5.3

Method	Dice $\uparrow$		NSD $\uparrow$		HD95 $\downarrow$	
	S	M	S	M	S	M
MedPLIB-7B [†]	47.2	36.0	37.1	27.3	220.7	378.2
MedPixel-7B	41.1	47.4	26.8	33.5	152.3	123.0
MedPixel-3B	40.5	47.3	26.2	33.2	163.5	132.5
BiomedParse	35.8	44.0	25.6	32.4	224.3	177.1
UniBiomed-1B	31.8	36.9	18.6	22.8	230.6	229.2
VisionReasoner-7B	26.5	33.3	12.0	18.0	291.7	243.6
LISA-7B	19.7	20.1	7.3	8.9	404.1	432.4
LISA++-7B	17.8	16.6	4.5	4.7	433.2	490.7
MMedAgent-7B	13.7	12.2	5.1	5.5	419.2	449.3
SAM4MLLM-8B	11.9	12.9	2.5	4.2	559.1	580.0
PixelLM-7B	11.1	10.4	1.9	1.9	539.7	567.4

Table 2: Zero-shot transfer on MeCoVQA-G+. S/M denote slice-level micro and modality-level macro averages. Bold denotes the best zero-shot result. [†] marks a benchmark-trained reference.

Dice at the slice level and +3.4 Dice at the modality level. It also reduces HD95 from 224.3/177.1 to 152.3/123.0, indicating more stable boundary localization across modalities.

U-MRG-14K. U-MRG-14K is a medical reasoning grounding benchmark introduced by MedReasoner (Yan et al. 2026), containing 2,480 test samples across ten modalities. Each sample pairs an implicit clinical query with a target mask. MedPixel is evaluated directly on the test set without benchmark-specific training. Results for the remaining

General MLLMs		Medical MLLMs		Grounding MLLMs	
Method	Dice	Method	Dice	Method	Dice
GPT-4o	4.72	MedR1-2B	14.73	VLMR1-REC-3B	22.19
Gemini-2.5-flash	14.29	MiniInternVL-4B	4.76	SegZero-7B	26.05
Qwen2.5-VL-7B	22.73	MedGemma-4B	8.90	SAM4MLLM-8B	16.49
InternVL3-8B	9.23	HuatuoGPT-7B	19.76	MedReasoner-7B	37.78
Qwen2.5-VL-72B	29.71	Lingshu-7B	16.48	MedPixel-7B	37.91
InternVL3-78B	7.23	Chiron-o1-8B	10.05		

Table 3: Zero-shot transfer on U-MRG-14K.

Scale	SFT	PLPO	T2 Reasoning			T4 Explanatory		
			Dice $\uparrow$	NSD $\uparrow$	MET $\uparrow$	Dice $\uparrow$	NSD $\uparrow$	MET $\uparrow$
3B	✓		58.2	46.5	50.5	72.1	51.2	39.7
	✓	✓	62.7	49.5	50.2	73.9	52.2	40.4
7B	✓		62.9	49.8	51.0	75.6	54.5	41.5
	✓	✓	66.7	53.4	51.3	76.5	55.2	42.3

Table 4: Effect of Pixel-Level Preference Optimization (PLPO) on T2 and T4. MET. denotes METEOR.

methods in Table 3 are taken from the published MedReasoner evaluation, where MLLMs generate spatial prompts for a fixed MedSAM2 segmenter. MedPixel-7B achieves comparable performance to MedReasoner-7B (37.91 vs. 37.78 Dice), while producing masks within a single unified model. This result further demonstrates that the resulting unified pixel-language model transfers effectively to unseen reasoning-grounding benchmarks.

4.4 Ablation Study

We present the main findings here and defer further analyses to Appendix E.

Effect of PLPO. Table 4 shows that PLPO consistently improves Dice and NSD on T2 and T4 at both model scales, with larger gains on reasoning segmentation, while METEOR remains broadly stable. Since the segmentation projector and mask decoder are frozen during Stage 2, these improvements indicate better alignment between generated `<SEG>` representations and pixel-level outcomes.

Dice-Reasoning Alignment. We use gpt-5.6-sol to determine whether each reasoning trace identifies the ground-truth target and compare these labels with Dice-based preferences. Reasoning-correct responses achieve much higher mean Dice than incorrect ones, both before PLPO (84 vs. 34) and after PLPO (87 vs. 42). Among preference pairs, 83.6% favor a reasoning-correct response over an incorrect one, while only 0.4% show the reverse, demonstrating strong alignment between Dice-based preferences and target-level reasoning.

Robustness to Box Perturbations. To evaluate robustness to imperfect box prompts, we scale the width and height of each GT-tight box by a factor s while keeping its center fixed and clipping the resulting box to the image boundaries. Here, $s = 1$ denotes the exact box, whereas $s < 1$ and $s > 1$ produce tighter and looser prompts, respectively. Figure 3 shows that SAM specialists perform strongly with exact boxes but degrade markedly as the box scale deviates from

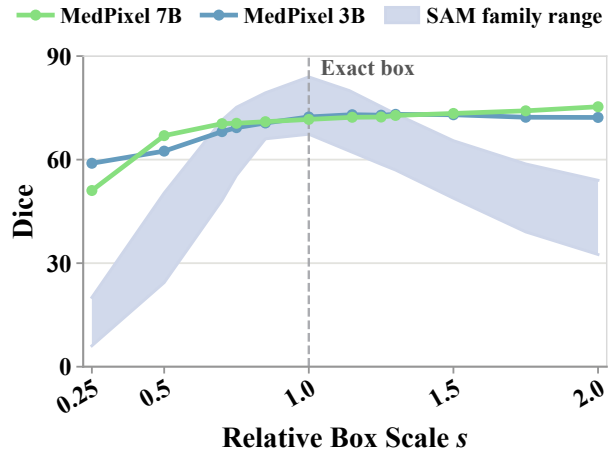


Figure 3: Robustness to box scaling on T3. Box width and height are scaled by s with the center fixed; $s = 1$ denotes the GT-tight box. The shaded region denotes the performance range of SAM-family baselines.

1. In contrast, MedPixel remains stable over a broad range of scales, demonstrating greater robustness to approximate spatial guidance.

5 Limitations

MedPixel has several limitations. First, because MedPLG-440K is constructed by repurposing existing segmentation annotations, its coverage is limited to the anatomical structures and imaging modalities represented in the source datasets. Although the synthesis process requires neither additional annotation for individual samples nor external LLM annotation, language generated from rules and templates is less diverse than natural clinical interactions. Second, PLPO requires ground-truth masks for offline preference mining, limiting its applicability when pixel-level supervision is unavailable. Third, our evaluation mainly covers 2D or slice-based images and single-turn interactions, leaving volumetric, longitudinal, and multi-turn pixel-language reasoning underexplored. Finally, reference-based language metrics cannot fully assess the clinical correctness and completeness of generated explanations, calling for further evaluation by clinical experts.

6 Conclusion

This work presents **MedPixel**, a unified model for medical language understanding, reasoning, interaction, and pixel-level localization. **MedPLG-440K** converts existing segmentation annotations into diverse pixel-language supervision, while **PLPO** uses mask quality to improve the alignment between generated responses and segmentation outcomes. Across five tasks, MedPixel achieves strong performance, with particularly clear gains on reasoning and explanatory segmentation. It also generalizes well to external benchmarks and remains robust to imperfect spatial prompts. Taken together, these results demonstrate the feasibility of unifying medical language reasoning, interaction, and pixel-

level localization within a single framework, with model conclusions grounded in precise visual evidence.

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
- Bannur, S.; Bouzid, K.; Castro, D. C.; Schwaighofer, A.; Thieme, A.; Bond-Taylor, S.; Ilse, M.; Pérez-García, F.; Salvatelli, V.; Sharma, H.; et al. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.
- Chen, H.; Fang, Z.; Singla, Y.; and Dredze, M. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3563–3599.
- Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024a. HuatuoGPT-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.
- Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; et al. 2024b. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 7346–7370.
- Chen, Y.-C.; Li, W.-H.; Sun, C.; Wang, Y.-C. F.; and Chen, C.-S. 2024c. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, 323–340. Springer.
- Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sam-med2d. *arXiv preprint arXiv:2308.16184*.
- Deng, Z.; He, R.; Liu, J.; Wang, Y.; Meng, Z.; Jiang, S.; Xie, Y.; and Liu, Z. 2025. Med-GLIP: Advancing medical language-image pre-training with large-scale grounded dataset. *arXiv preprint arXiv:2508.10528*.
- FreedomIntelligence. 2025. Medical-R1-Distill-Data. <https://huggingface.co/datasets/FreedomIntelligence/Medical-R1-Distill-Data>. SFT data distilled from DeepSeek-R1 on HuatuoGPT-o1 verifiable medical problems.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2): 3.
- Huang, X.; Shen, L.; Liu, J.; Shang, F.; Li, H.; Huang, H.; and Yang, Y. 2025a. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3779–3787.
- Huang, Y.; Peng, Z.; Zhao, Y.; Yang, P.; Yang, X.; and Shen, W. 2025b. Medseg-r: Reasoning segmentation in medical images with multimodal large language models. *arXiv preprint arXiv:2506.10465*.
- Huang, Y.; Yang, X.; Zhou, H.; Cao, Y.; Dou, H.; Dong, F.; and Ni, D. 2024. Robust box prompt based sam for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 1–11. Springer.
- Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.
- Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2567–2577.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.
- Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9579–9589.
- Li, B.; Yan, T.; Pan, Y.; Luo, J.; Ji, R.; Ding, J.; Xu, Z.; Liu, S.; Dong, H.; Lin, Z.; et al. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8745–8760.
- Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.
- Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 1650–1654. IEEE.
- Liu, Y.; Ma, Z.; Pu, J.; Qi, Z.; Wu, Y.; Shan, Y.; and Chen, C. 2026. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *Advances in Neural Information Processing Systems*, 38: 126078–126108.
- Liu, Y.; Qu, T.; Zhong, Z.; Peng, B.; Liu, S.; Yu, B.; and Jia, J. 2025. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv e-prints*, arXiv:2505.
- Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature communications*, 15(1): 654.
- Ma, J.; Yang, Z.; Kim, S.; Chen, B.; Baharoon, M.; Fallahpour, A.; Asakereh, R.; Lyu, H.; and Wang, B. 2025. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.

- Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.
- Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2025. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, 28085–28128.
- Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.
- Rückert, J.; Bloch, L.; Brüngel, R.; Idrissi-Yaghir, A.; Schäfer, H.; Schmidt, C. S.; Koitka, S.; Pelka, O.; Abacha, A. B.; G. Seco de Herrera, A.; et al. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1): 688.
- Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Wang, G.; Gao, M.; Yang, S.; Zhang, Y.; He, L.; Huang, L.; Xiao, H.; Zhang, Y.; Li, W.; Chen, L.; et al. 2025a. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*.
- Wang, G.; Zhao, J.; Liu, X.; Liu, Y.; Cao, X.; Li, C.; Liu, Z.; Sun, Q.; Zhou, F.; Xing, H.; et al. 2025b. Citrus-V: Advancing Medical Foundation Models with Unified Medical Image Grounding for Clinical Reasoning. *arXiv preprint arXiv:2509.19090*.
- Wu, L.; Nie, Y.; He, S.; Zhuang, J.; Luo, L.; Li, T.; Xie, Z.; Chen, D.; Zhao, Y.; Mahboobani, N.; et al. 2025. Unibiomed: A universal foundation model for grounded biomedical image interpretation. *arXiv preprint arXiv:2504.21336*.
- Yan, Z.; Diao, M.; Yang, Y.; Jing, R.; Xu, J.; Zhang, K.; Yang, L.; Liu, Y.; Liang, K.; and Ma, Z. 2026. Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 11577–11585.
- Yang, S.; Qu, T.; Lai, X.; Tian, Z.; Peng, B.; Liu, S.; and Jia, J. 2023. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*.
- Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Zhao, T.; Gu, Y.; Yang, J.; Usuyama, N.; Lee, H. H.; Naumann, T.; Gao, J.; Crabtree, A.; Abel, J.; Moung-Wen, C.; et al. 2024. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*.

Appendix

Supplementary Material Overview

Appendix A	Data Sources
Appendix B	MedPLG-440K Construction
Appendix C	Training Data and Optimization
Appendix D	Evaluation Protocols
Appendix E	Additional Experimental Results
Appendix F	Qualitative Results
Appendix G	Licenses

A Data Sources

This section summarizes the data used to construct MedPLG-440K and the auxiliary corpora used during Stage 1 training. Detailed dataset construction and training-mixture composition are provided in Appendices B and C.

A.1 Segmentation Data Sources

MedPLG-440K is constructed from the training annotations aggregated by BiomedParse (Zhao et al. 2024). Table 5 lists the corresponding original datasets according to the subset mapping provided by BiomedParse.

A.2 Medical QA and Clinical-Reasoning Sources

In addition to MedPLG-440K, Stage 1 incorporates image-based medical VQA data from PubMedVision, ROCov2, PMC-VQA, and SLAKE (Chen et al. 2024b; Rückert et al. 2024; Zhang et al. 2023; Liu et al. 2021); text-based QA data from PubMedQA, MedMCQA, MedQA, and Medbullets (Jin et al. 2019; Pal, Umapathi, and Sankarasubbu 2022; Jin et al. 2021; Chen et al. 2025); and clinical-reasoning data from HuatuoGPT-o1, Medical-R1-Distill, and Citrus-S3 (Chen et al. 2024a; FreedomIntelligence 2025; Wang et al. 2025a). PubMedVision is additionally used to synthesize image-grounded reasoning samples. Table 6 summarizes these auxiliary sources.

B MedPLG-440K Construction

MedPLG-440K contains 444,297 pixel–language records constructed from existing medical segmentation annotations across four tasks: T1 Referring, T2 Reasoning, T3 Interactive, and T4 Explanatory. The construction uses deterministic image–mask measurements, curated clinical knowledge, and predefined templates, without external LLM annotation.

B.1 Dataset Composition

Each source sample provides a medical image, its ground-truth mask, the target category, and imaging modality. Visual attributes are first extracted from the image–mask pair and then combined with task-specific templates to generate four interaction formats.

T1 directly refers to a named target, T2 requires target inference from indirect clinical clues, T3 introduces a point or box prompt, and T4 combines segmentation with a medical explanation. A source annotation may be reused across multiple formats. Table 7 summarizes the resulting composition.

Task	Records	Primary supervision
T1 Referring	94,261	Explicit target reference
T2 Reasoning	21,367	Indirect clinical query
T3 Interactive	318,670	Point or box guidance
T4 Explanatory	9,999	Grounded medical explanation
Total	444,297	

Table 7: Composition of MedPLG-440K.

B.2 Visual Attribute Extraction

For each image–mask pair, nine numerical measurements are deterministically converted into seven language descriptors: location, size, shape, components, boundary, contrast, and uniformity. Let $I \in \mathbb{R}^{H \times W}$ denote the image, $M \in \{0, 1\}^{H \times W}$ the target mask, and $A = |M|$ its foreground area.

Location. The mask centroid (c_x, c_y) is assigned to a 3×3 spatial grid using boundaries at $W/3$, $2W/3$, $H/3$, and $2H/3$. This produces central, left, right, upper, lower, and four corner descriptors.

Size. The relative area

$$r_A = \frac{A}{HW} \quad (12)$$

is mapped to tiny, small, moderate-sized, large, or extensive using cutoffs at 0.02, 0.08, 0.20, and 0.45.

Shape. For tight bounding-box dimensions (w_b, h_b) , we compute

$$\rho = \frac{\max(w_b, h_b)}{\min(w_b, h_b)}, \quad \eta = \frac{A}{w_b h_b}, \quad (13)$$

where ρ is the aspect ratio and η the foreground fill ratio. The target is classified as irregular if $\eta < 0.45$; otherwise, it is round if $\rho < 1.3$, oval if $\rho < 2.0$, and elongated otherwise.

Components. Let K denote the number of connected foreground components. We use single region for $K \leq 1$, a few scattered foci for $2 \leq K \leq 3$, and multiple scattered foci for $K \geq 4$.

Boundary. From the largest external contour, we compute

$$C = \frac{4\pi A_c}{P^2}, \quad V = \frac{A_c}{A_{\text{hull}}}, \quad R = \frac{P}{\sqrt{A_c}}, \quad (14)$$

where A_c , P , and A_{hull} denote the contour area, perimeter, and convex-hull area. The boundary score is

$$s = 2\mathbb{I}(C > 0.75) + \mathbb{I}(0.55 < C \leq 0.75) \\ + 2\mathbb{I}(V > 0.90) + \mathbb{I}(0.75 < V \leq 0.90) \\ + 2\mathbb{I}(R < 4.5) + \mathbb{I}(4.5 \leq R < 6.5), \quad (15)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Scores $s \geq 5$, $s = 4$, $s = 3$, $s = 2$, and $s < 2$ correspond to well-circumscribed and smooth margins, well-defined margins, partially defined margins, lobulated and partially irregular margins, and irregular and ill-defined margins, respectively.

Subset	Original dataset	Subset	Original dataset
ACDC	Automated Cardiac Diagnosis Challenge	LGG	LGG MRI Segmentation
amos22	AMOS 2022	LIDC-IDRI	LIDC-IDRI
BreastUS	BUSI	LiverUS	Ultrasound Simulation and Segmentation
CAMUS	CAMUS	MMs	M&Ms
CDD-CESM	CDD-CESM	MSD	Medical Segmentation Decathlon
COVID-19_CT	COVID-19 CT Segmentation Dataset	NeoPolyp	BKAI-IGH NeoPolyp
COVID-QU-Ex	COVID-QU-Ex	OCT-CME	Intraretinal Cystoid Fluid Dataset
CXR_Masks_and Labels	Chest Xray Masks and Labels	PanNuke	PanNuke
DRIVE	DRIVE	PolypGen	PolypGen
FH-PS-AOP	FH-PS-AOP	QaTa-COV19	QaTa-COV19
G1020	G1020	Radiography	COVID-19 Radiography Database
GlaS	GlaS	REFUGE	REFUGE
ISIC	ISIC 2018	siim-acr- pneumothorax	SIIM-ACR Pneumothorax
kits23	KiTS23	UWaterloo SkinCancer	UWaterloo Skin Cancer Dataset

Table 5: Segmentation datasets underlying MedPLG-440K. Subset identifiers and dataset mappings follow the BiomedParse organization.

Data group	Dataset
Image VQA	PubMedVision (Jin et al. 2019)
	ROCOv2 (Rückert et al. 2024)
	PMC-VQA (Zhang et al. 2023)
	SLAKE (Liu et al. 2021)
Text QA	PubMedQA (Jin et al. 2019)
	MedMCQA (Pal, Umapathi, and Sankarasubbu 2022)
	MedQA (Jin et al. 2021)
	Medbullets (Chen et al. 2025)
Clinical reasoning	HuatuoGPT-o1 (Chen et al. 2024a)
	Medical-R1-Distill (FreedomIntelligence 2025)
	Citrus-S3 (Wang et al. 2025a)

Table 6: Medical question-answering and reasoning datasets used during Stage 1 training.

Contrast. The foreground-to-background intensity ratio

$$I_r = \frac{\mu_{fg}}{\mu_{bg}} \quad (16)$$

is classified as high for $I_r > 1.15$, low for $I_r < 0.85$, and iso otherwise. These levels are expressed using the modality-specific terms in Table 8.

Uniformity. The foreground intensity standard deviation σ_{fg} is mapped to homogeneous for $\sigma_{fg} < 12$, relatively uniform for $\sigma_{fg} < 30$, mildly heterogeneous for $\sigma_{fg} < 55$, and markedly heterogeneous otherwise. For homogeneous regions, the wording is adapted to the modality: internal echotexture for ultrasound, attenuation for CT, density for mammography, signal intensity for MRI, and internal texture otherwise.

Modality	High	Iso	Low
CT	hyperdense	isodense	hypodense
MRI	hyperintense	isointense	hypointense
Ultrasound	hyperechoic	isoechoic	hypoechoic
X-ray	radiopaque	—	radiolucent
Mammography	hyperdense	isodense	hypodense
OCT	hyperreflective	—	hyporefective

Table 8: Modality-specific terminology for relative intensity.

B.3 Task-Specific Synthesis

The extracted visual information is incorporated into predefined task-specific templates together with category names, imaging modalities, clinical clues, and spatial prompts. T1 directly uses the category name, T2 and T4 use the complete descriptor set, and T3 uses only location, size, and shape. Unavailable contrast and uniformity clauses are omitted automatically.

T1 Referring T1 explicitly names the target category. Its prompt bank covers seven styles, including direct commands, polite requests, questions, technical expressions, and presence checks. A representative prompt is “Segment the {category}.” The response is sampled from five short templates, such as “Sure, the segmentation mask is <seg>.”

T2 Reasoning T2 omits the category name and describes the target through functional, anatomical, symptomatic, diagnostic, exclusion-based, or appearance-based clues. A representative prompt is “Segment the structure responsible for {function}.” The response contains a template-constructed reasoning trace, a conclusion identifying the inferred target, and an observation instantiated from the extracted descriptors:

```
<think> {opening} → {clue interpretation} → {category inference} → {segmentation approach} </think>
```

Task	Instruction and response
T1 Referring	User: Segment the liver. Assistant: The target is < seg >.
T2 Reasoning	User: Where is the large solid organ in the right upper quadrant of the abdomen? Segment it. Assistant: <think> This requires anatomical reasoning to determine the segmentation target. The query specifies the large solid organ in the right upper quadrant of the abdomen as the defining landmark. The liver is the structure that fits the description of the large solid organ in the right upper quadrant of the abdomen. I will now produce a precise mask for the liver. </think> This is the liver, located as described. I can see the liver in the central region. It is a large, oval structure displaying well-circumscribed, smooth margins, hyperechoic relative to surrounding tissue, with mildly heterogeneous texture. < seg >
T3 Interactive	User: Find the edges of the organ here < ref > and tell me how big it is. Assistant: This is the liver in abdominal ultrasound. It presents in the central region as a large, oval structure. < seg >
T4 Explanatory	User: Provide a clinical assessment of this ultrasound and segment the relevant structures. Assistant: <think> My task is to identify the liver in this ultrasound based on its imaging characteristics. I can distinguish the liver from adjacent structures based on its typical anatomical position and imaging appearance. The segmentation will follow the margin of the liver, capturing its complete extent in this view. </think> On careful review, the liver is visible in the central region as a large, oval structure with well-circumscribed, smooth margins, hyperechoic relative to surrounding tissue, with mildly heterogeneous texture. It receives dual blood supply from the hepatic artery and portal vein. < seg >

Table 9: Representative MedPLG-440K examples.

{conclusion} The {category} occupies the {location} region as a {size} {components} with {shape} contours and {boundary}{contrast clause}{uniformity clause}. <|seg|>

T3 Interactive T3 pairs an instruction with either a positive foreground point or a tight bounding box. Point and box prompts are sampled from separate template banks, for example, “Segment the foreground object marked by the point.” For non-organ categories, organ is replaced with structure. The response identifies the indicated target and describes its location, size, and shape:

This structure is the {category}, situated in the {location} region with a {size}, {shape} appearance. <|seg|>

T4 Explanatory T4 uses explicit, implicit, and clinical prompt families with probabilities 0.40, 0.30, and 0.30, respectively. Explicit prompts name the target, implicit prompts request pathology detection, and clinical prompts request general image analysis. Implicit prompts fall back to the explicit form for non-pathological targets.

The response combines a template-constructed reasoning trace, a visual observation, and an optional medical knowledge sentence:

<think> {opening} → {visual analysis} → {segmentation approach} → {optional modality cue} </think>
{prefix} The {category} is visible in the {location} region as a {size}, {shape} structure with {components} and {boundary}{contrast clause}{uniformity clause}. {medical knowledge} <|seg|>

Template diversity is further increased through predefined clinical phrases, synonym substitution for T2 clues, modality-specific terminology, and automatic article correction. Table 9 presents verbatim examples generated from a shared liver annotation.

Data source	Instances	Share
MedPLG-440K grounding	591,291	32%
Medical VQA and QA	1,178,760	64%
Distilled clinical reasoning	72,751	4%
Total	1,842,802	100%

Table 10: Effective Stage 1 training mixture.

C Training Data and Optimization

C.1 Stage 1 Training Mixture

Stage 1 combines MedPLG-440K, public medical VQA and QA corpora, and distilled clinical-reasoning data. Table 10 summarizes the effective mixture after task-level replication.

T1 and T4 are repeated twice, T2 three times, and T3 once, yielding

$$\begin{aligned}
 N_{\text{ground}} &= 2 \times 94,261 + 3 \times 21,367 \\
 &\quad + 318,670 + 2 \times 9,999 \\
 &= 591,291.
 \end{aligned} \tag{17}$$

Replication changes only sampling frequency; MedPLG-440K retains 444,297 unique records.

C.2 Stage 1 Supervised Fine-Tuning

The 3B and 7B variants use the same Stage 1 training configuration. For the joint objective defined in the main paper, we set $\lambda_{\text{seg}} = 1$, $\lambda_{\text{focal}} = 100$, and $\lambda_{\text{dice}} = \lambda_{\text{iou}} = \lambda_{\text{obj}} = 5$. The focal loss uses $\alpha = 0.25$ and $\gamma = 2$. Samples without pixel-level annotations are optimized only with the language modeling loss.

The LoRA dropout is set to 0.1. The language- and vision-side LoRA adapters, multimodal and task-specific projection modules, token embeddings, and language output head use a learning rate of 1.5×10^{-5} . The SAM2 image encoder, prompt encoder, and mask decoder are jointly optimized with a smaller learning rate of 5×10^{-6} .

Training uses AdamW for one epoch with a cosine learning-rate schedule and a 3% warmup ratio. We use a per-device batch size of 1, gradient accumulation over 4 steps, a maximum sequence length of 4,096, and bf16 precision.

C.3 Stage 2 Pixel-Level Preference Optimization

Candidate generation. For each T2 reasoning-segmentation input, the Stage 1 policy generates $N = 8$ candidate responses: one greedy decode, three stochastic samples with temperature 0.3 and top- $p = 0.9$, and four samples with temperature 0.7 and top- $p = 0.95$. The maximum generation length is 1,024 new tokens.

Pair construction. Following the filtering and scoring procedure described in the main paper, we set $\tau_{\min} = 0.30$ and $\tau_{\text{gap}} = 0.10$. When multiple lower-scoring candidates satisfy the required Dice margin, the highest-scoring one is retained as the rejected response. Inputs without a qualified pair are discarded, resulting in 1,786 preference pairs from the T2 training branch.

Optimization settings. The Stage 1 checkpoint initializes the trainable policy and serves as the frozen reference policy. Reference log-probabilities are precomputed and cached before training. We set $\beta = 0.25$ and $\lambda_{\text{nl}} = 0.05$, without response-length normalization.

Stage 2 is trained for three epochs with a learning rate of 1×10^{-6} . The per-device batch size is 1 with gradient accumulation over 4 steps. The maximum DPO sequence length is 2,048, and the maximum gradient norm is 0.5.

The original language and visual backbone weights remain frozen. Stage 2 optimizes their $q/k/v/o$ LoRA adapters, together with the visual merger, token embeddings, and language output head. The segmentation projector and the complete SAM2 branch remain frozen.

All training is conducted on NVIDIA RTX PRO 6000 GPU.

D Evaluation Protocols

D.1 In-Domain Evaluation Sets

T1–T4 are constructed from the official BiomedParse test directories, without introducing an additional train–test split. Sequence construction is performed independently within each official split.

Clip construction. Eight volumetric datasets, including ACDC, MSD, AMOS22, KiTS23, LGG, LIDC-IDRI, MMs, and COVID-19 CT, are treated as 3D data. Their slices are grouped by volume identity and target category, ordered by slice index, and divided when the gap between adjacent annotated slices is at least three. Sequences longer than 16 frames are further divided using a window length of 16 and a stride of 8. Intermediate slices without target annotations are

retained as negative frames with empty masks. Images from the remaining 2D datasets are represented as single-frame clips.

This procedure produces 24,391 test clips, comprising 17,489 single-frame clips and 6,902 multi-frame clips.

T1 evaluation. T1 evaluates all 24,391 test clips using one frame per clip. For a clip with L frames, the frame at index $\lfloor L/2 \rfloor$ is selected. If the middle frame does not contain a valid target mask, an annotated frame from the same clip is used instead.

T2–T4 evaluation. T2, T3, and T4 are sampled from the official test pool using dataset-balanced sampling with a fixed random seed of 123. Candidates must be single-frame samples with valid masks, and duplicate frame–mask pairs are removed. T2 further restricts targets to the predefined organ-knowledge set. The resulting sets contain 2,500 T2 samples, 2,000 T3 samples, and 2,500 T4 samples. T3 contains equal numbers of point-prompted and box-prompted examples.

T5 evaluation. T5 contains 10,267 multiple-choice questions from public medical QA benchmarks, including 1,647 image-based and 8,620 text-based samples.

D.2 External Evaluation

We evaluate zero-shot transfer on the official test sets of MeCoVQA-G+ and U-MRG-14K, containing 2,719 and 2,480 evaluation samples, respectively. No benchmark-specific fine-tuning or preference optimization is performed.

For MeCoVQA-G+, slice-level results are aggregated over all samples to obtain micro scores, while modality-level scores are computed separately and averaged to obtain macro performance. U-MRG-14K is evaluated using its official segmentation protocol.

D.3 Evaluation Metrics

For a predicted mask $\hat{M}$ and ground-truth mask M , Dice is computed as

$$\text{Dice}(\hat{M}, M) = \frac{2|\hat{M} \cap M|}{|\hat{M}| + |M|}. \quad (18)$$

Normalized surface Dice is computed with a boundary tolerance of five pixels. HD95 measures the 95th-percentile bidirectional surface distance, with lower values indicating better boundary agreement.

For T2 and T4, METEOR (Banerjee and Lavie 2005) measures response similarity, while target grounding accuracy measures whether the normalized target concept is fully recovered. Empty or invalid responses and outputs without a decodable mask receive zero scores.

For T5, the predicted option is extracted from the generated response and evaluated using multiple-choice accuracy.

E Additional Experimental Results

E.1 Detailed Interactive Segmentation Results

T3 interactive segmentation contains two equally sized subsets: 1,000 point-prompted samples and 1,000 box-prompted

Prompt	MedPixel-3B		MedPixel-7B	
	Dice $\uparrow$	NSD $\uparrow$	Dice $\uparrow$	NSD $\uparrow$
Box	72.36	49.39	71.65	49.79
Point	79.65	54.88	80.16	56.47
Avg.	76.00	52.14	75.91	53.13

Table 11: Detailed T3 interactive-segmentation results under point and box prompts. Each subset contains 1,000 samples. Avg. denotes the arithmetic mean over the two prompt types.

samples. While the main paper reports the average performance over the two subsets, Table 11 presents the separate Dice and NSD results.

As shown in Table 11, point prompts consistently outperform box prompts for both model scales. For MedPixel-3B, point prompting improves Dice and NSD by 7.29 and 5.49 points, respectively, while MedPixel-7B achieves gains of 8.51 Dice and 6.68 NSD. MedPixel-7B obtains the strongest point-prompted performance, reaching 80.16 Dice and 56.47 NSD.

The stronger performance under point prompts is also encouraging for practical use, as a point can be specified with a single click and requires less interaction effort than drawing a bounding box. These results suggest that MedPixel can achieve accurate interactive segmentation from lightweight spatial guidance.

E.2 Detailed Effect of Pixel-Level Preference Optimization

We provide a detailed comparison between the Stage 1 SFT and Stage 2 PLPO checkpoints across all five tasks. For T1–T4, we report Dice, gIoU, cIoU, and NSD. For the language-generative T2 and T4 tasks, we additionally report METEOR (MET.) and target grounding accuracy (TGA). TGA measures whether all tokens in the reference target phrase are recovered in the generated response. For T5, IMCQ and TMCQ denote image-based and text-based medical multiple-choice accuracy, respectively. Higher values are better for all metrics. Dashes indicate that the corresponding NSD results were not retained during the earlier 3B Stage 1 evaluation.

Table 12 shows that PLPO produces its largest improvements on T2 reasoning segmentation. For MedPixel-3B, Dice, gIoU, and cIoU increase by 4.51, 4.10, and 5.58 points, respectively. The corresponding gains for MedPixel-7B are 3.75, 3.37, and 4.79 points, together with a 3.55-point improvement in NSD. These consistent changes show that PLPO improves both sample-level overlap and dataset-level pixel aggregation.

T4 explanatory segmentation also benefits from PLPO. For MedPixel-3B, Dice, gIoU, and cIoU improve by 1.81, 1.58, and 2.01 points, respectively. MedPixel-7B obtains smaller but consistent improvements across all four segmentation metrics. METEOR and TGA also improve on T4 for both model scales, indicating that the pixel-level gains do not degrade the generated explanations.

On T2, the response metrics remain broadly stable. ME-

TEOR changes by only -0.3 and $+0.3$ points for the 3B and 7B models, respectively, while TGA changes by $+1.2$ and -1.7 points. These variations are small relative to the segmentation gains, suggesting that PLPO mainly improves the pixel-level outcomes while preserving the learned reasoning behavior.

In contrast, T1 referring segmentation and T3 interactive segmentation remain nearly unchanged after PLPO. Their Dice, gIoU, cIoU, and NSD values exhibit only minor fluctuations, with no consistent degradation. T5 medical VQA is similarly preserved: IMCQ improves at both scales, while TMCQ remains nearly unchanged for MedPixel-3B and improves slightly for MedPixel-7B.

Overall, PLPO selectively improves reasoning- and explanation-oriented segmentation while maintaining referring segmentation, interactive segmentation, and medical question-answering performance.

E.3 Dice–Reasoning Alignment

We conduct an additional analysis to examine whether the pixel-level preference signal used by PLPO is consistent with target-level reasoning. The analysis covers the 2,500 responses generated on the T2 test set before and after PLPO, together with the 1,786 mined preference pairs. The headline comparisons are reported in the main paper; here, we describe the reasoning-annotation protocol and provide complementary statistics.

Reasoning annotation protocol. We use gpt-5.6-sol to determine whether each generated `<think>` trace identifies the ground-truth segmentation target. For each response, the judge receives the segmentation query, the ground-truth target, and the generated reasoning trace. The image, predicted mask, and Dice score are excluded from the judging input, so the reasoning label is assigned independently of segmentation quality.

The judge returns one of three labels: *yes*, *partial*, or *no*. A response is labeled *yes* when its concluded target matches the ground-truth structure, allowing synonyms, clinically equivalent expressions, and equivalent image- and patient-side laterality descriptions. A *partial* label denotes the correct organ with an incorrect substructure or a genuinely ambiguous conclusion. A *no* label denotes a different structure, incorrect laterality, or the absence of an identifiable target. Only *yes* is counted as reasoning-correct; *partial*, *no*, and unparseable outputs are treated as incorrect. The same fixed rubric is applied to the SFT and PLPO outputs and to both responses in every preference pair.

The exact judging prompt is reproduced below.

System prompt. You are a strict judge of medical image-segmentation reasoning. For each item you get: QUESTION (what to segment), GT (the ground-truth target structure whose mask is the correct answer), and REASONING (the model’s `<think>` text). Decide whether the REASONING concludes a target that is the SAME structure as GT. Allow synonyms, laterality equivalence between image-side and patient-side wording, and clinical equivalence. Judge IDENTITY ONLY against GT; ignore mask and segmentation quality, and ignore whether the wording in QUESTION appears to contradict GT, since GT is authoritative.

Scale	Training	T1 Referring				T2 Reasoning					T3 Interactive				T4 Explanatory					T5 VQA			
		Dice	gIoU	cloU	NSD	Dice	gIoU	cloU	NSD	MET.	TGA	Dice	gIoU	cloU	NSD	Dice	gIoU	cloU	NSD	MET.	TGA	IMCQ	TMCQ
3B	Stage 1	84.29	76.95	82.15	60.39	58.19	50.99	47.92	46.45	50.5	62.9	75.72	67.94	72.69	52.16	72.12	65.35	67.06	51.19	39.7	82.4	42.14	52.46
	Stage 2	84.30	76.96	82.15	60.40	62.70	55.09	53.50	49.46	50.2	64.1	76.00	68.17	72.90	52.14	73.93	66.93	69.07	52.18	40.4	83.4	42.87	52.45
7B	Stage 1	85.02	77.83	83.00	61.65	62.90	55.83	53.48	49.80	51.0	72.5	75.97	68.84	74.34	53.22	75.63	68.59	71.28	54.50	41.5	87.6	47.30	57.84
	Stage 2	85.00	77.81	82.97	61.67	66.65	59.20	58.27	53.35	51.3	70.8	75.91	68.83	73.75	53.13	76.46	69.42	71.96	55.24	42.3	87.9	47.48	58.14

Table 12: Detailed effect of Stage 2 Pixel-Level Preference Optimization across all five tasks. MET. denotes METEOR, and TGA denotes target grounding accuracy.

Label `yes` when the concluded target equals GT or an equivalent structure; label `partial` when the correct organ is identified but the substructure is incorrect, or when the conclusion is genuinely ambiguous; and label `no` when the response concludes a different structure, uses incorrect laterality, or contains no clear target. Return only a compact JSON array, with one object per item and no additional prose: `{"uid": "...", "c": "yes|no|partial", "t": "concluded target"}`.

User-message template. Judge these items:
`{"uid": "...", "question": "<segmentation query>", "gt": "<ground-truth target>", "reasoning": "<generated reasoning trace>"}`.

Response-level association. As shown in Fig. 4(a), reasoning-correct responses decode substantially better masks under both policies. Their mean Dice scores are 84.4 versus 33.8 for SFT and 86.6 versus 41.7 after PLPO. Treating Dice as a score for distinguishing reasoning-correct from reasoning-incorrect responses yields AUC values of 0.861 and 0.842, respectively. The consistently high AUC indicates that the association between target-level reasoning and mask quality remains strong after preference optimization.

Preference-level association. Figure 4(b) shows that 93.5% of chosen responses are reasoning-correct, compared with 10.3% of rejected responses. At the pair level, 83.6% prefer a reasoning-correct response over an incorrect one, whereas only 0.4% exhibit the reverse ordering. Among pairs whose candidates differ in reasoning correctness, the higher-Dice response is correct in 99.5% of cases. Thus, the preference ordering induced by mask quality is highly consistent with target-level reasoning.

Reasoning preservation after PLPO. On the same 2,500 T2 test samples, PLPO increases mean Dice from 62.9 to 66.7, while the reasoning-correct rate changes from 57.4% to 55.6%. This difference is not statistically significant under McNemar’s test ($p = 0.07$). Moreover, reasoning changes from correct to incorrect while Dice improves in only 1.4% of samples. These results indicate that PLPO improves segmentation without a detectable systematic degradation in target-level reasoning.

E.4 Detailed Evaluation of Generated Responses

Evaluation protocol. We further evaluate the pixel-level predictions and generated responses on the 2,500-sample T2 reasoning-segmentation and T4 explanatory-segmentation test sets. For T2, we retain the complete reasoning and final answer after removing structural tags, segmentation tokens, coordinate tokens, and chat templates. For T4, we evaluate the final explanatory text after the reasoning block.

Evaluation metrics. Dice and NSD evaluate region overlap and boundary agreement, respectively. Avg. Len. reports the average number of content words after text cleaning. METEOR (MET.) measures similarity to the reference response. Target grounding accuracy (TGA) requires all tokens in the reference target phrase to be recovered, while Head Match requires recovery of its principal anatomical or finding term. Avg. Len. has no preferred direction.

Results. Table 13 shows that MedPixel achieves the strongest overall combination of pixel-level prediction and language generation on both tasks. On T2, MedPixel-7B obtains the best Dice, NSD, METEOR, TGA, and Head Match results. VisionReasoner is the strongest baseline with complete reasoning responses, but it remains substantially behind MedPixel in both segmentation quality and implicit target recovery. LISA++, PixelLM, and UniPixel mainly produce short templates or segmentation-oriented outputs, resulting in limited target recovery. Citrus-V is included for pixel-level comparison, but its per-sample T2 reasoning responses are unavailable.

On T4, MedPixel-7B again achieves the highest Dice, NSD, and METEOR, together with high target-recovery accuracy. Citrus-V obtains perfect TGA and Head Match because its short responses almost always contain the target name. However, its substantially lower METEOR and segmentation scores indicate that recovering the target name alone does not constitute a complete and accurately grounded explanation. VisionReasoner produces much longer responses but remains behind MedPixel in mask quality, reference similarity, and target recovery.

Target recovery is generally more difficult on T2 because its query does not explicitly name the target. The model must first infer the intended anatomical structure or finding before producing the mask and reasoning response. T4 instead evaluates the description of an identified target, leading to higher target-recovery scores and shorter responses. Overall, MedPixel consistently combines accurate masks with complete and target-consistent language generation across both reasoning and explanatory segmentation.

E.5 Robustness to Bounding-Box Perturbations

Experimental setup. We use the same 1,000-sample box test set as in the T3 evaluation, with a fixed random seed of 42. All methods receive identically perturbed boxes and are evaluated against the original ground-truth masks at their native resolution. We report Dice and normalized surface Dice (NSD) with a tolerance of five pixels.

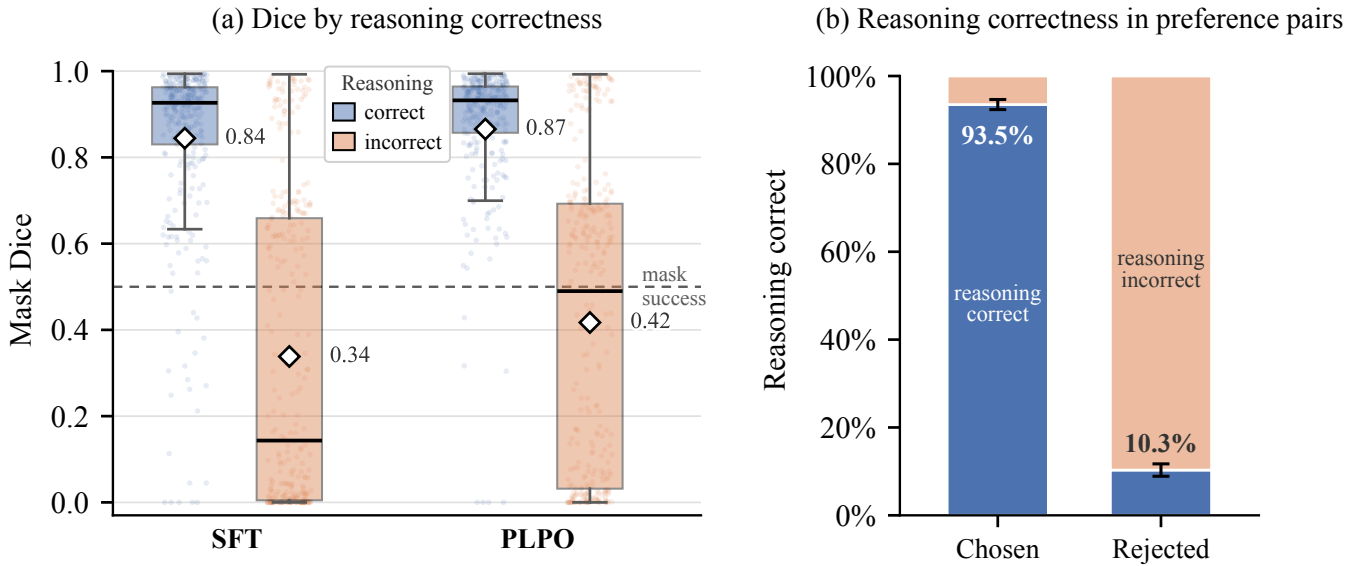


Figure 4: Dice-reasoning alignment on T2. (a) Reasoning-correct responses achieve substantially higher Dice before and after PLPO. (b) Higher-Dice chosen responses are substantially more likely to be reasoning-correct than rejected responses.

Method	T2 Reasoning Segmentation						T4 Explanatory Segmentation					
	Dice↑	NSD↑	Len.	MET.↑	TGA↑	Head↑	Dice↑	NSD↑	Len.	MET.↑	TGA↑	Head↑
Citrus-V-8B	37.5	25.9	—	—	—	—	35.1	23.0	16	10.2	100.0	100.0
VisionReasoner-7B	33.4	14.5	87	29.6	39.8	47.4	31.1	13.0	182	21.2	72.0	76.9
UniBiomed-1.4B	28.3	14.4	86	4.9	27.3	29.9	35.8	19.1	55	7.3	23.7	25.9
LISA++-7B	22.8	8.9	6	7.9	0.1	0.2	12.1	5.3	46	12.0	9.3	11.1
PixelLM-7B	18.2	3.0	19	8.1	4.9	5.8	10.0	1.5	30	11.6	14.2	15.1
UniPixel-7B	27.9	7.5	2	3.0	0.0	0.0	30.2	9.5	5	3.2	0.0	0.0
MedPixel-3B	62.7	49.5	92	50.2	64.1	73.9	73.9	52.2	44	40.4	83.4	87.8
MedPixel-7B	66.7	53.4	92	51.3	70.8	79.2	76.5	55.2	42	42.3	87.9	91.0

Table 13: Detailed evaluation of pixel-level predictions and generated responses on T2 reasoning segmentation and T4 explanatory segmentation. Avg. Len. is measured in content words, and TGA denotes target grounding accuracy based on recovery of all target tokens.

Scale-only perturbation. To isolate the effect of box size, we keep the original box center fixed and symmetrically scale its width and height. Let (c_x, c_y) , w , and h denote the original box center, width, and height, respectively. The scaled box is defined as

$$c'_x = c_x, \quad c'_y = c_y, \quad w' = sw, \quad h' = sh. \quad (19)$$

Here, $s = 1.00$ preserves the original tight box, while $s < 1.00$ shrinks it and $s > 1.00$ enlarges it.

As shown in Table 14 and Figure 6, the SAM-family methods are strongly dependent on box scale. Their performance generally peaks near the original tight box at $s = 1.00$ and decreases when the box is either contracted or enlarged. By contrast, MedPixel exhibits a substantially flatter performance profile.

MedPixel is particularly stable for enlarged boxes. MedPixel-7B increases from 71.65 Dice at $s = 1.00$ to 75.32 at $s = 2.00$, whereas all SAM-family baselines degrade as

additional background is introduced. Stronger degradation is observed under severe contraction, because the box may no longer contain the complete target.

Joint shift-and-scale perturbation. Starting from the tight ground-truth bounding box, we randomly shift its center in the horizontal and vertical directions, and independently contract or expand its left, right, top, and bottom sides relative to the shifted center. The resulting box therefore contains both positional displacement and asymmetric extent variation.

Let the original box have center (c_x, c_y) , width w , height h , and center-to-side distances d_k , where $k \in \{l, r, t, b\}$. We perturb the box as

$$\begin{aligned} c'_x &= c_x + \epsilon_x w, & \epsilon_x &\sim \mathcal{U}(-r, r), \\ c'_y &= c_y + \epsilon_y h, & \epsilon_y &\sim \mathcal{U}(-r, r), \\ d'_k &= \alpha_k d_k, & \alpha_k &\sim \mathcal{U}(1 - r, 1 + r). \end{aligned} \quad (20)$$

Method	$s = 0.25$	$s = 0.50$	$s = 0.70$	$s = 0.75$	$s = 0.85$	$s = 1.00$	$s = 1.15$	$s = 1.25$	$s = 1.30$	$s = 1.50$	$s = 1.75$	$s = 2.00$
MedSAM	6.06	24.37	48.20	55.42	70.80	83.97	80.01	75.38	73.26	63.86	53.14	44.59
MedSAM2	20.08	50.35	71.45	75.18	79.36	80.88	77.98	74.45	72.55	65.48	58.76	54.02
SAM2	16.99	42.92	65.33	70.21	76.44	77.21	70.29	64.26	60.46	48.82	39.05	32.51
SAM-Med2D	13.27	34.03	54.12	58.92	66.02	67.36	62.49	58.66	56.98	51.06	44.33	39.67
MedPixel-3B	58.94	62.46	68.11	69.26	70.62	72.36	73.02	72.88	73.12	73.00	72.28	72.22
MedPixel-7B	51.05	66.95	70.42	70.53	70.99	71.65	72.27	72.34	72.80	73.38	74.14	75.32

Table 14: Dice under scale-only bounding-box perturbations. The box center remains fixed, while its width and height are multiplied by s ; $s = 1.00$ corresponds to the original tight box.

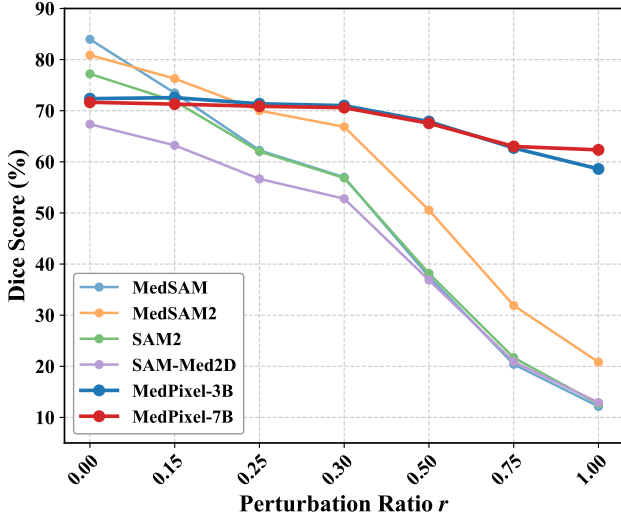


Figure 5: Dice under joint bounding-box shift-and-scale perturbations. Larger r indicates stronger center displacement and side variation. MedPixel degrades more gradually than the SAM-family baselines.

The four side-scaling factors are sampled independently. The resulting box is

$$B' = (c'_x - d'_l, c'_y - d'_t, c'_x + d'_r, c'_y + d'_b). \quad (21)$$

Here, $r = 0.00$ retains the original tight box, while larger values produce stronger center displacement and side variation.

As shown in Table 15 and Figure 5, the SAM-family methods perform strongly with tight boxes but degrade rapidly as the perturbation increases. Their Dice curves decline sharply once the box is substantially displaced or asymmetrically resized. In contrast, both MedPixel variants remain comparatively stable across increasing perturbation strengths.

At $r = 1.00$, MedPixel-7B retains 62.34 Dice and 43.26 NSD, corresponding to decreases of only 9.31 and 6.53 points, respectively. The comparatively flat MedPixel curves indicate that its predictions are less dependent on precise agreement between the input box and the target extent.

Qualitative analysis. We further examine representative endoscopic-polyp and chest-X-ray examples under both perturbation settings. In the joint-perturbation visualizations,

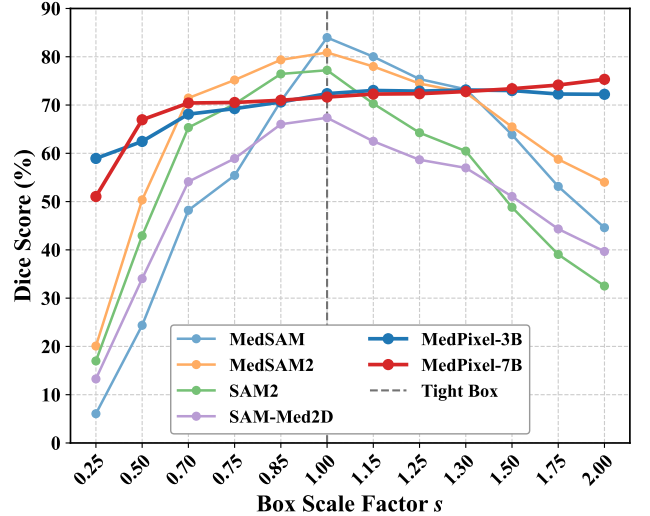


Figure 6: Dice under scale-only bounding-box perturbations. The dashed line marks the tight box at $s = 1.00$. MedPixel remains comparatively stable as the box is contracted or enlarged.

green, magenta, and yellow indicate the ground-truth contour, predicted contour, and perturbed box, respectively, and D denotes the Dice score of the displayed prediction.

Scale-only examples. Figures 7 and 8 visualize the effect of changing the box size while preserving its center. The two examples illustrate the different effects of box contraction and enlargement across target scales.

The scale-only examples explain the asymmetric behavior observed in Table 14. Enlarged boxes preserve the complete target and mainly introduce additional background, which MedPixel can largely suppress. In contrast, strongly contracted boxes may remove essential target regions and provide insufficient spatial evidence. This effect is especially visible for the large bilateral lung target, where MedPixel-7B is better able to recover the complete anatomy from an incomplete box. For the smaller and visually distinctive polyp, both MedPixel variants remain stable over a broad range of box scales.

As a geometric reference, using the tight box directly as the foreground mask yields 66.18 Dice and 15.44 NSD, con-

Method	$r = 0.00$		$r = 0.15$		$r = 0.25$		$r = 0.30$		$r = 0.50$		$r = 0.75$		$r = 1.00$		$\Delta_{0 \rightarrow 1}$	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
MedSAM	83.97	62.10	73.47	41.82	62.25	31.43	56.98	27.96	37.56	19.04	20.39	12.70	12.14	8.20	-71.83	-53.90
MedSAM2	80.88	57.14	76.28	48.07	70.02	39.38	66.85	35.79	50.55	25.02	31.90	17.28	20.82	12.06	-60.06	-45.08
SAM2	77.21	41.47	71.87	34.83	62.02	27.99	56.84	25.71	38.18	17.68	21.70	11.99	12.64	7.72	-64.57	-33.75
SAM-Med2D	67.36	36.17	63.23	31.69	56.66	26.83	52.79	24.60	36.87	18.07	20.90	12.28	12.90	8.57	-54.46	-27.60
MedPixel-3B	72.36	49.39	72.56	49.60	71.36	48.45	71.00	48.02	67.89	45.09	62.71	41.15	58.61	38.95	-13.75	-10.44
MedPixel-7B	71.65	49.79	71.29	49.46	70.86	49.15	70.60	48.82	67.54	46.66	63.01	43.46	62.34	43.26	-9.31	-6.53

Table 15: Robustness to joint bounding-box shift-and-scale perturbations. The perturbation ratio r controls both center displacement and independent variation of the four box sides, with $r = 0.00$ corresponding to the original tight box.

firming that MedPixel predicts target boundaries rather than simply reproducing the box. Together, the quantitative and qualitative results show that MedPixel can use image and target evidence to remain stable under inaccurate spatial prompts, with degradation occurring primarily when severe perturbations remove a substantial portion of the target.

Joint-perturbation examples. Figures 9 and 10 provide representative comparisons for a localized endoscopic lesion and a large bilateral anatomical structure.

The qualitative examples support the aggregate trend in Table 15. For the localized polyp, MedPixel continues to recover the target contour even when the perturbed box is visibly displaced or asymmetrically resized. The lung example reveals a more difficult failure mode: when the box excludes a substantial portion of the bilateral anatomy, the prediction may become incomplete. Nevertheless, MedPixel-7B remains stable over a wider perturbation range, whereas the SAM-family methods frequently fragment, expand into background regions, or collapse to an incorrect structure.

F Qualitative Results

Figures 11–15 present representative outputs across the five MedPixel task types. For T1–T4, each example shows the user prompt, input image, predicted mask, and ground-truth mask. T2 and T4 additionally display the generated reasoning and final grounded response, while T3 demonstrates interaction through point or bounding-box prompts. T5 presents medical question-answering examples that do not require pixel-level prediction.

Across the five task types, MedPixel adapts its output to the requested interaction form. T1 directly grounds an explicitly named target, whereas T2 infers the intended structure from functional or clinical evidence before localization. T3 follows point- and box-based spatial prompts, and T4 couples visual explanation with mask prediction. T5 further demonstrates that the unified model preserves medical question-answering capability when no pixel-level output is required.

F.1 Licenses

MedPLG-440K is derived from datasets aggregated by BiomedParse (Zhao et al. 2024). These source datasets are distributed under different licenses and usage conditions, and downstream users remain responsible for complying with the terms specified by the original data providers. We do not redistribute the underlying medical images. MedPLG-440K

and MedPixel are intended for research use only and have not been validated for clinical diagnosis or treatment decisions.

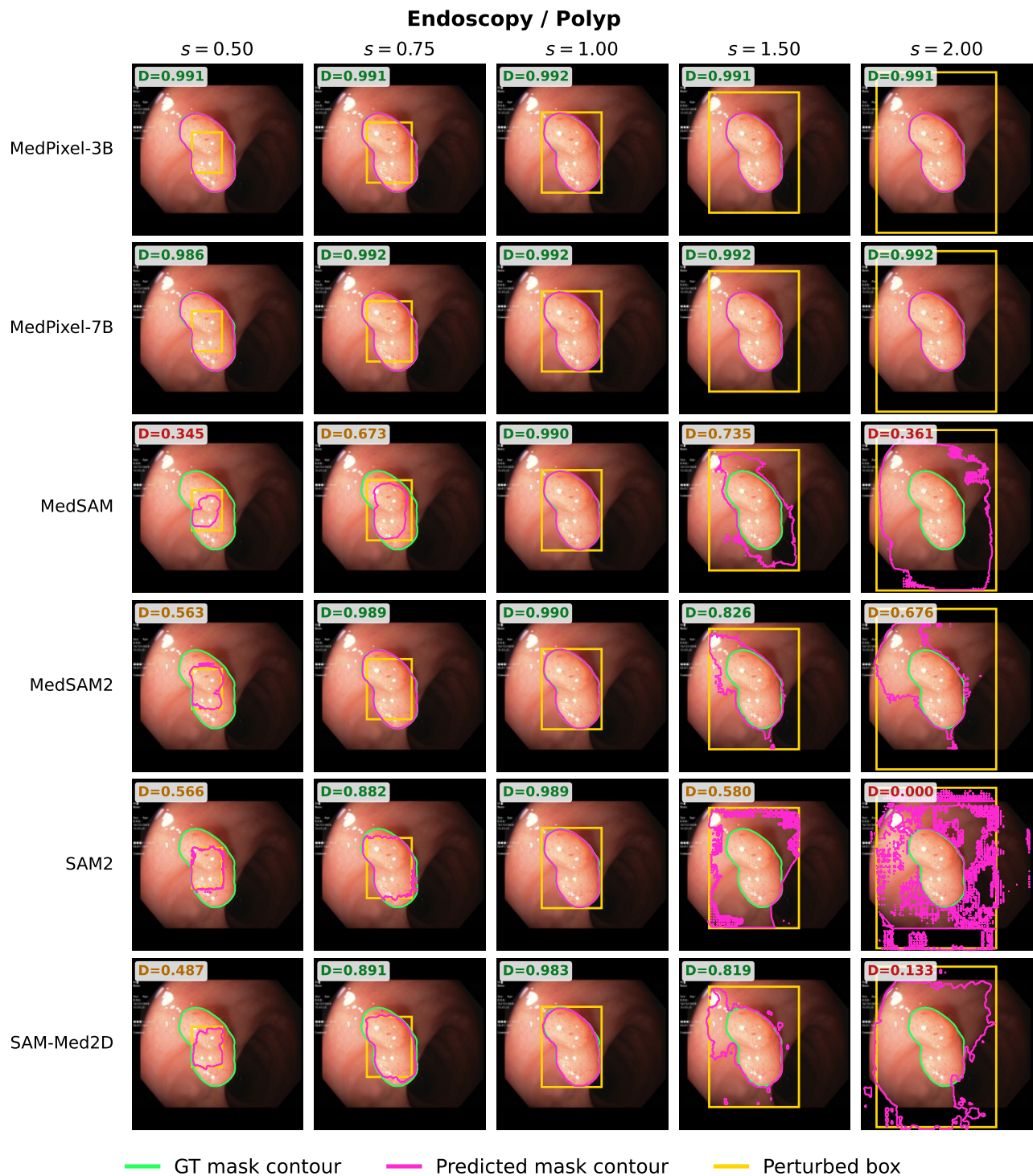


Figure 7: Qualitative robustness to scale-only perturbations for an endoscopic polyp. MedPixel preserves an accurate target contour across both contracted and enlarged boxes, while the SAM-family methods are more sensitive to deviations from the original tight box.

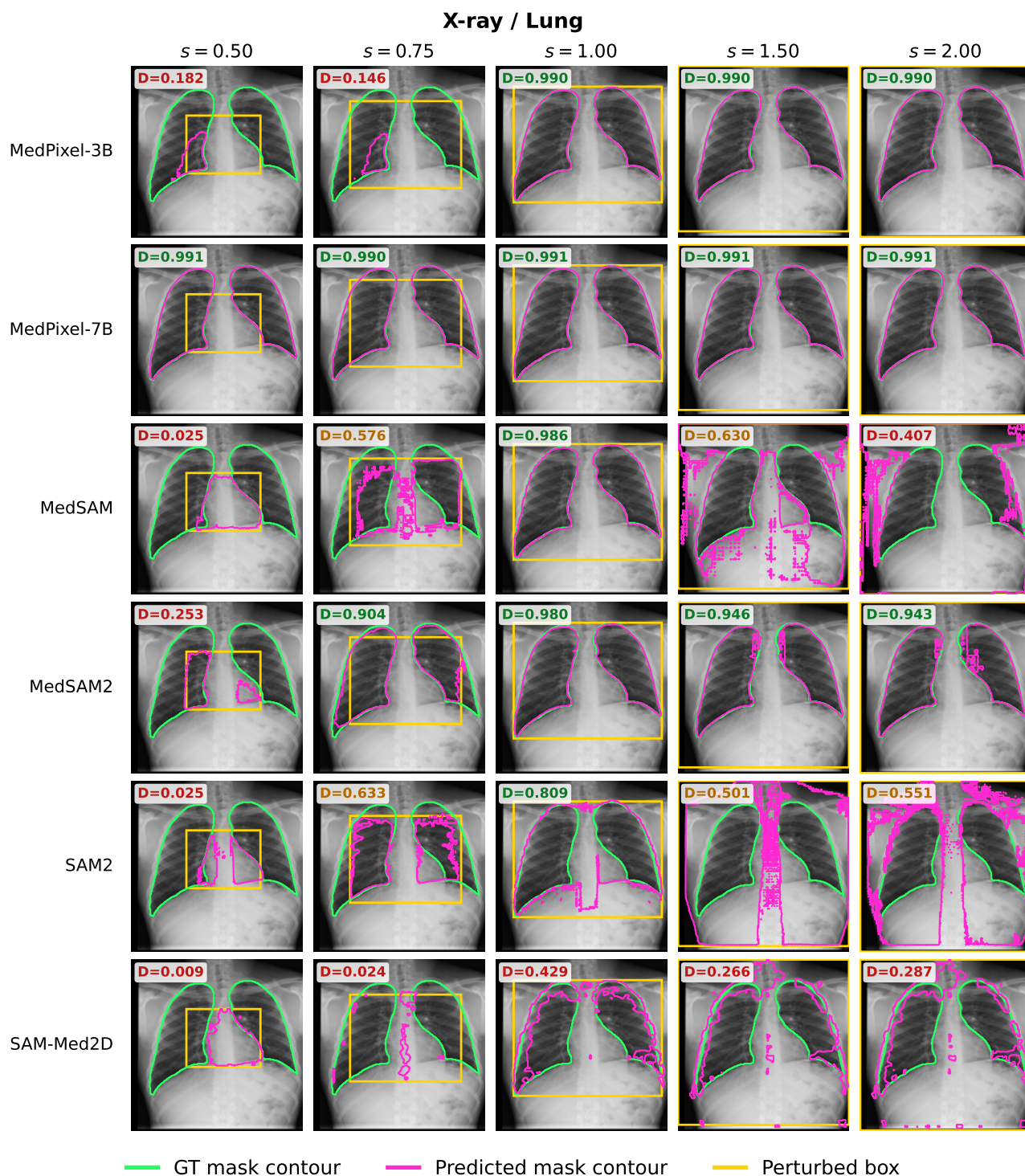


Figure 8: Qualitative robustness to scale-only perturbations for the lungs in a chest X-ray. Severe contraction removes a substantial portion of the target and is therefore more challenging than box enlargement. MedPixel-7B remains stable across the evaluated scales, whereas MedPixel-3B is more sensitive when the box covers only a small central part of the lungs.

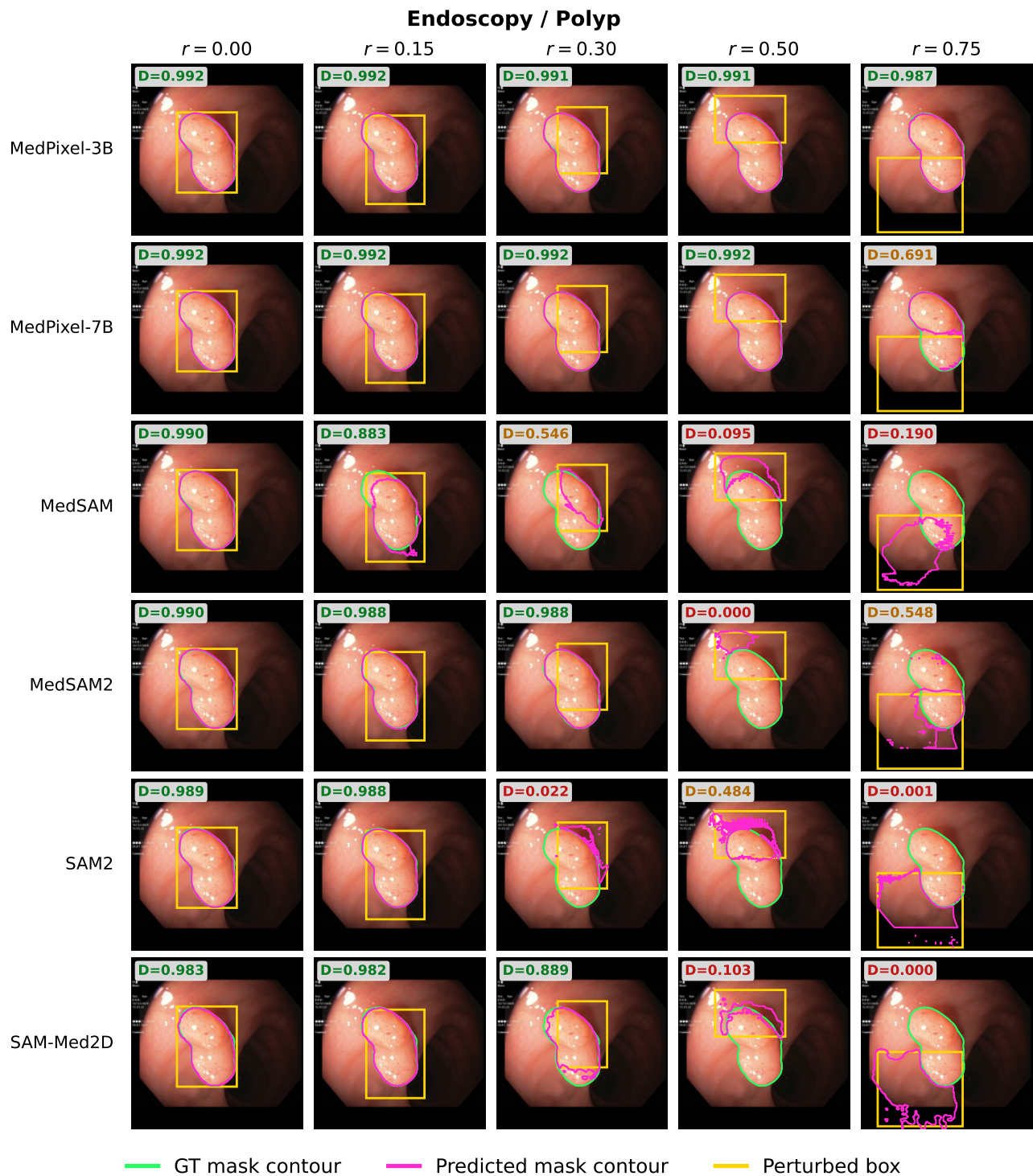


Figure 9: Qualitative robustness to joint shift-and-scale perturbations for an endoscopic polyp. The perturbation strength increases from left to right. MedPixel remains closely aligned with the target under moderate box displacement and asymmetric resizing, whereas the compared SAM-family methods become increasingly sensitive to the perturbed box.

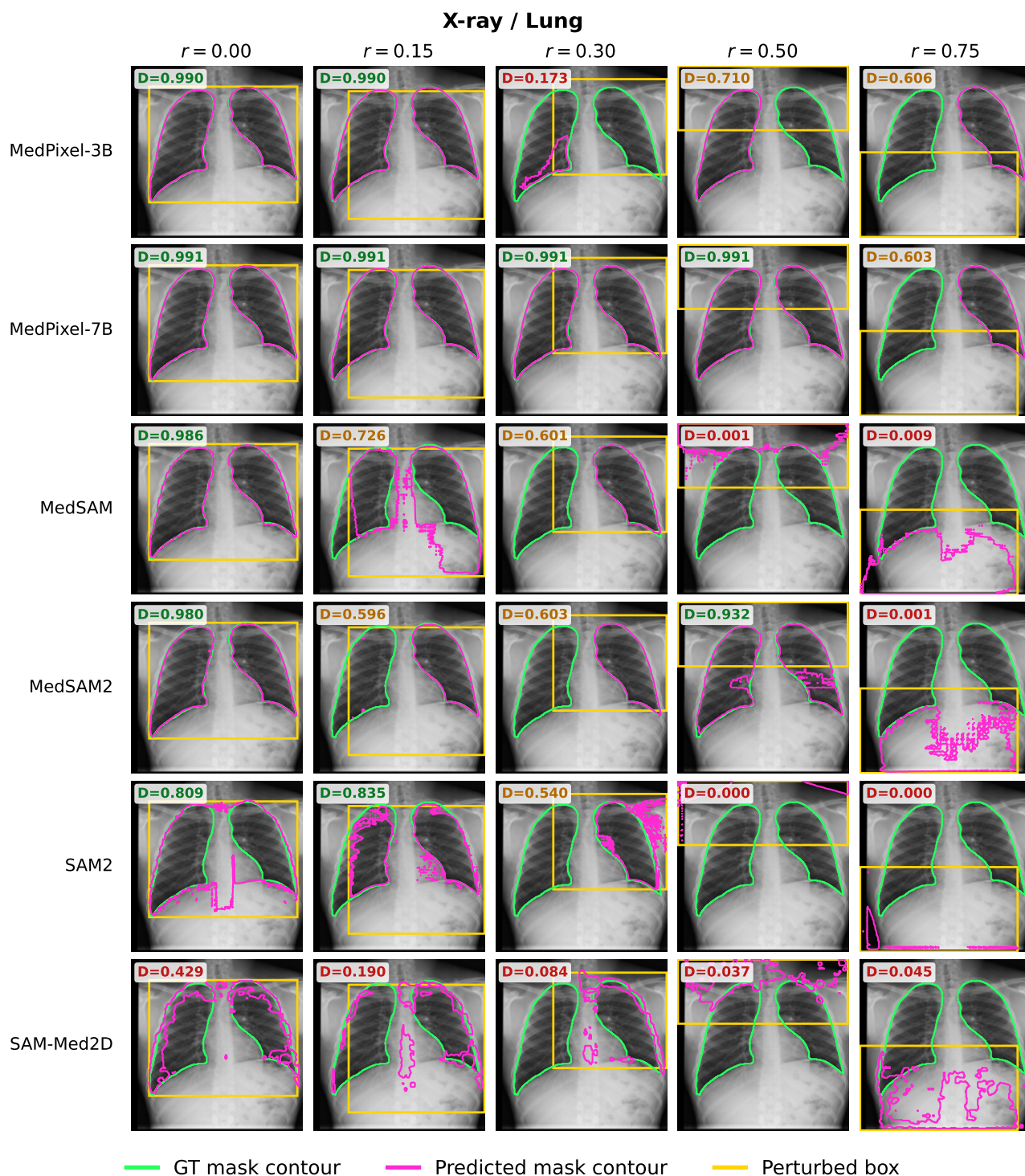
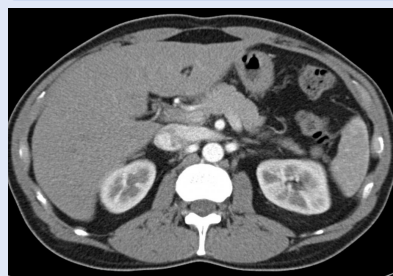


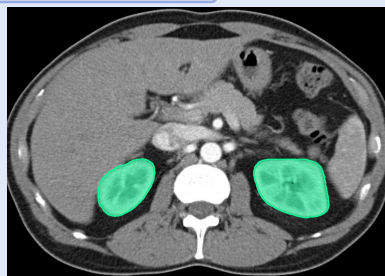
Figure 10: Qualitative robustness to joint shift-and-scale perturbations for the lungs in a chest X-ray. The lung example is more challenging because a displaced box may exclude a large portion of the bilateral target. MedPixel-7B remains accurate under moderate perturbations, while severe target exclusion eventually causes degradation.

T1 Referring Segmentation

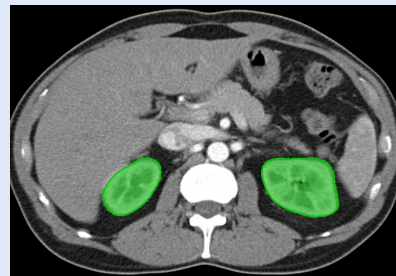
User: Please segment kidney in abdominal CT.



Input Image

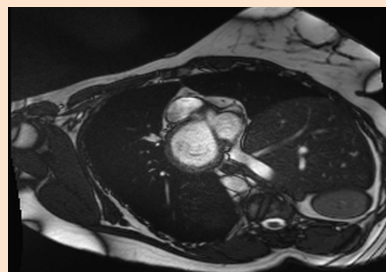


Pred Mask

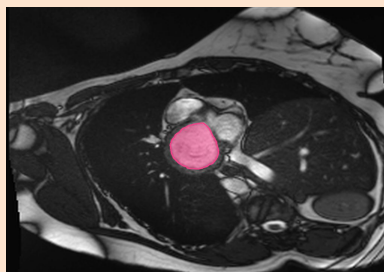


Ground Truth

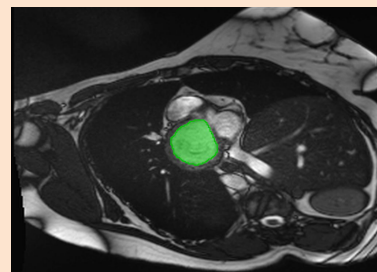
User: Please segment left heart ventricle.



Input Image

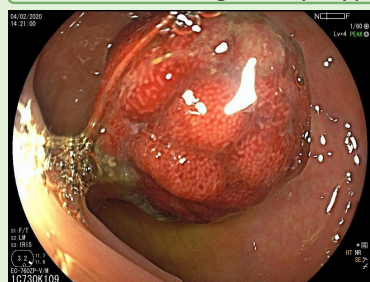


Pred Mask

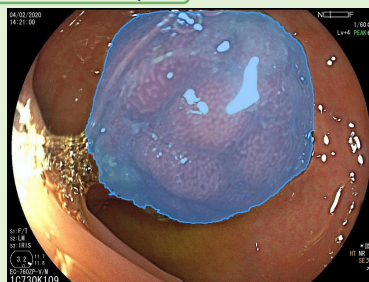


Ground Truth

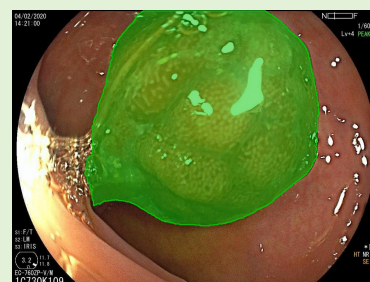
User: Please segment polyp in colon endoscope.



Input Image



Pred Mask

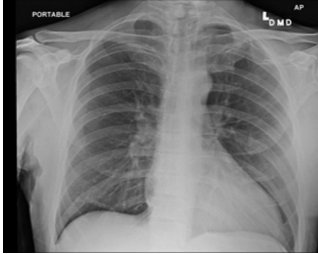


Ground Truth

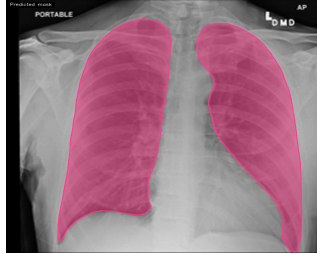
Figure 11: **Qualitative results for T1 referring segmentation.** Given an explicitly specified target, MedPixel directly predicts the corresponding mask. The examples cover kidney segmentation in abdominal CT, left-ventricle segmentation in cardiac MRI, and polyp segmentation in colonoscopy.

T2 Reasoning Segmentation

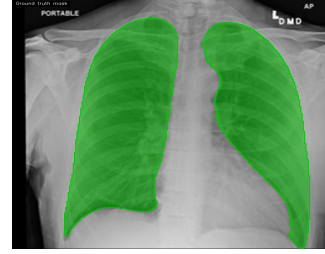
User: Segment the structure responsible for respiratory ventilation and perfusion throughout the thoracic cavity.



Input Image



Pred Mask



Ground Truth

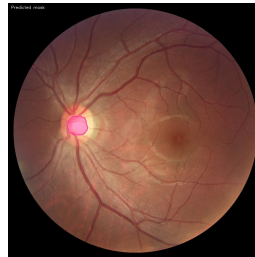
Thinking: I need to reason about the target based on the given description. The query references the organ responsible for respiratory ventilation and perfusion throughout the thoracic cavity. The lung is well-known for its role in respiratory ventilation and perfusion throughout the thoracic cavity. I will now produce a precise mask for the lung.

Answer: The lung is confirmed as the structure involved in respiratory ventilation and perfusion throughout the thoracic cavity. The lung is visualized in the central region, appearing large with round form, partially defined margins, radiolucent relative to surrounding tissue, with mildly heterogeneous texture. <|seg|>.

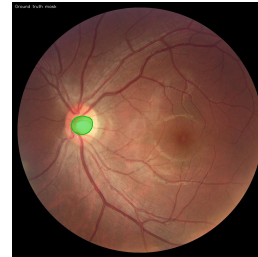
User: The findings are consistent with glaucoma. Provide the segmentation mask.



Input Image



Pred Mask



Ground Truth

Thinking: I need to identify the target based on the clinical clue. The radiological pattern is consistent with glaucoma. The diagnosis of glaucoma involves the eye, which is the expected site. I will segment the optic cup based on this reasoning.

Answer: The eye is the site of glaucoma in this case. The optic cup is located in the left region, demonstrating tiny dimensions, round contour, and well-circumscribed, smooth margins, with relatively uniform parenchymal texture. <|seg|>.

Figure 12: **Qualitative results for T2 reasoning segmentation.** MedPixel infers the intended target from functional or clinical clues before producing the grounded response and segmentation mask. The examples identify the lungs from their physiological function and the optic cup from a glaucoma-related diagnostic clue.

T3 Interactive Segmentation

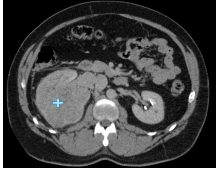
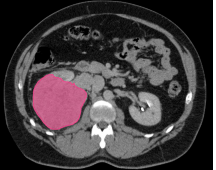
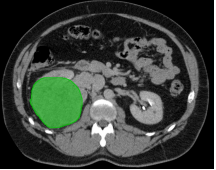
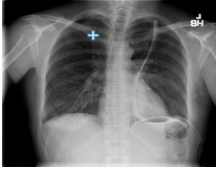
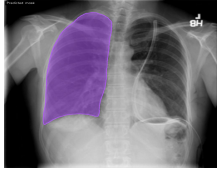
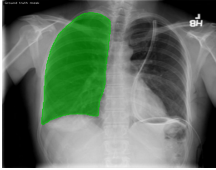
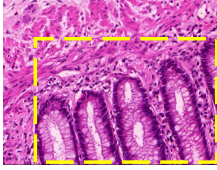
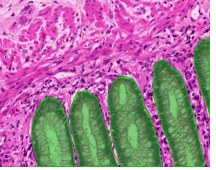
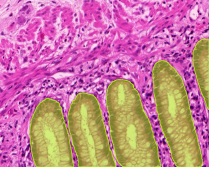
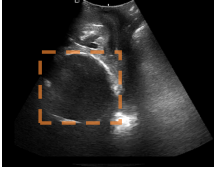
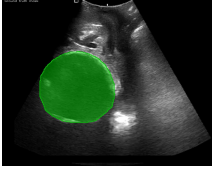
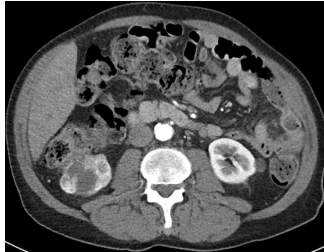
User: Can you segment the region I'm pointing to?			User: Segment the right lung.		
					
Input Image Pred Mask Ground Truth			Input Image Pred Mask Ground Truth		
Response: This is the kidney tumor in abdominal CT. It presents in the left region as a small, round structure. < seg >.			Response: This is the right lung in chest X-Ray, visible in the left region as a moderate-sized, oval structure. < seg >.		
User: Extract the foreground object confined within these bounding lines.			User: Segment the fetal head.		
					
Input Image Pred Mask Ground Truth			Input Image Pred Mask Ground Truth		
Response: The bounded area shows the glandular structure in colon pathology. It is located in the central region as a large, irregular structure. < seg >.			Response: The bounded area shows the fetal head in transperineal ultrasound. It is located in the central region as a moderate-sized, round structure. < seg >.		

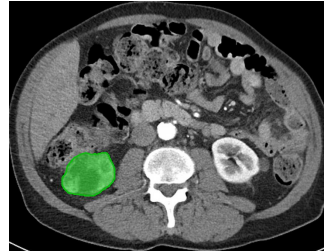
Figure 13: **Qualitative results for T3 interactive segmentation.** MedPixel identifies and segments targets indicated by point or bounding-box prompts. The examples cover abdominal CT, chest radiography, colon pathology, and transperineal ultrasound. The generic textual instruction specifies the requested operation, while the spatial prompt determines the target region.

T4 Explanatory Segmentation

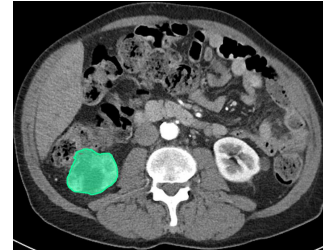
User: What is the kidney tumor in this image? Describe it and provide the segmentation mask.



Input Image



Pred Mask



Ground Truth

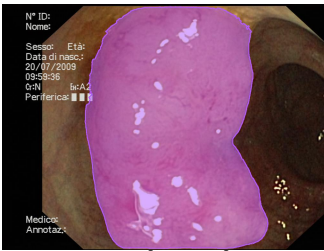
Thinking: I will analyze this CT scan to find the kidney tumor. The CT attenuation characteristics of the kidney tumor are consistent with its known tissue composition. The kidney tumor shows characteristic imaging features that differentiate it from the surrounding structures in this area. I will carefully outline the kidney tumor, following its contour to produce an accurate segmentation result.

Answer: A tiny, round kidney tumor is identified in the left region with well-circumscribed, smooth margins, hyperdense relative to surrounding tissue, with relatively uniform parenchymal texture. Notably, renal cell carcinoma is the most common primary renal malignancy, with an increased risk of metastasis to regional lymph nodes and distant sites. <|seg|>.

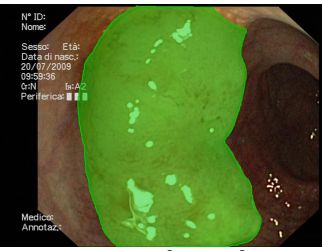
User: Can you describe the polyp in this endoscopy image and outline it with a segmentation mask?



Input Image



Pred Mask



Ground Truth

Thinking: I will analyze this endoscopy image to find the polyp. The tissue properties of the polyp, including its signal intensity and border characteristics, help identify it. I will produce a precise mask that follows the true border of the polyp based on the observed tissue boundaries.

Answer: Upon examination, a large polyp is present in the central region, displaying round margins, with mildly heterogeneous texture. Notably, polyp morphology on endoscopy includes pedunculated, sessile, and flat subtypes, each with different management implications. <|seg|>.

Figure 14: **Qualitative results for T4 explanatory segmentation.** MedPixel jointly generates a visual explanation and a segmentation mask. The responses identify and describe the location, morphology, boundary, and appearance of a kidney tumor in abdominal CT and a polyp in endoscopy.

T5 Medical VQA

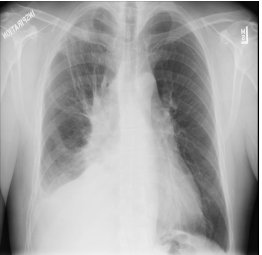
	Question: A 71-year-old male who had a left total hip arthroplasty a decade ago presents with progressive hip and thigh pain that began 18 months ago. In the last 6 weeks, the pain has become severe enough to prevent ambulation even with walker assistance. Physical examination reveals mild pain during passive internal and external hip rotation, and he is unable to walk during the office visit. Laboratory findings show: WBC 10,300, CRP 0.2, and ESR 13. Upon review of the provided radiograph (figure A), what is the most appropriate treatment strategy? A. Revision arthroplasty with a total femur prosthesis B. Revision arthroplasty with a fully coated cementless stem, cable wiring, and bone graft C. Revision arthroplasty with a modular, tapered stem and bone grafting of the diaphyseal fixation D. Radionuclide bone scan and MRI E. Open reduction internal fixation with a cable plate and allograft strut Response: B) Revision arthroplasty with a fully coated cementless stem, cable wiring, and bone graft ✓
	Question: What part of the body is the radiograph showing? A. Elbow B. Ankle C. Wrist D. Knee Response: C) Wrist ✓
	Question: How to prevent the occurrence of the disease on the lower right lobe of lung in this image? Response: The patient should avoid smoking and keep healthy. ✓

Figure 15: **Qualitative results for T5 medical VQA.** MedPixel retains general medical question-answering capability without requiring pixel-level output. The examples cover treatment-strategy selection, preventive guidance based on a chest radiograph, and anatomical recognition from radiographs.