SEGBANANA: STEERING UNIFIED MULTIMODAL MODELS INTO MEDICAL SEGMENTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation remains challenging in practical deployment, as models often struggle to generalize beyond the distributions covered by their training data and high-quality pixel-level annotations are typically unavailable for adaptation. Inspired by the cross-task transferability of large language models, we investigate whether unified multimodal models (UMMs) can transfer their pretrained visual understanding, reasoning, and generation capabilities to medical image segmentation without task-specific post-training. By recasting segmentation as structured visual generation, we find that frontier UMMs (e.g., Nano Banana) already exhibit basic segmentation capabilities across diverse clinical scenarios, but still struggle with challenging tasks requiring specialized anatomical or domain-specific knowledge. We further show that these limitations can be effectively mitigated by incorporating visual anatomical knowledge from in-context exemplars, expanding candidate solutions through repeated sampling, and refining suboptimal predictions via targeted editing. Motivated by these observations, we propose SegBanana, to our knowledge, the first agentic visual generation framework for training-free medical image segmentation. SegBanana builds on a frozen UMM as the core generative model, augmented with Anatomy-Aware Knowledge Retrieval and Comparative Quality Critique to unlock its potential segmentation capability. A State-Aware Multimodal Controller maintains structured state and iteratively orchestrates these tools, repeatedly refining intermediate predictions toward higher-quality masks. Across eight medical segmentation datasets, SegBanana achieves an average mDice of 77.45%, outperforming representative generalist (SAM3 and SegGPT) and medical-specific (BiomedParse and MedSAM3) baselines by at least 14.93 points, while remaining robust to out-of-domain visual supports. Ablations further confirm the effectiveness of agentic inference and the contributions of retrieval, critique, and structured state.

1 INTRODUCTION

Medical image segmentation remains difficult to deploy reliably in new clinical scenarios. Changes in imaging devices, acquisition protocols, clinical sites, patient populations, or disease presentations can induce substantial distribution shifts, under which most training-based models generalize poorly beyond their training distributions (Niu et al., 2024). Meanwhile, the pixel-level annotations required for adaptation are often scarce and costly to acquire at scale. These limitations motivate more flexible paradigms that can generalize to new clinical scenarios with limited task-specific supervision.

Recent efforts toward this objective have explored multiple directions, including medical-specific and generalist methods, as illustrated in Fig. 1. Medical-specific foundation models (Ma et al., 2024; Zhu et al., 2024; Zhao et al., 2025) and recent agentic methods (Jiang et al., 2026b; Huang et al., 2026) build on capabilities acquired from medical training data or pretrained segmentation tools, and often degrade on tasks or domains beyond this coverage. Generalist methods instead leverage broadly pretrained representations or annotated supports at inference time, but open-vocabulary methods (Barsellotti et al., 2025; Zeng et al., 2024; Carion et al., 2026) may lack specialized anatomical knowledge, while in-context methods (Cuttano et al., 2026; Zhang et al., 2024; 2023) rely heavily on well-matched supports and degrade under domain shift. Overall, existing methods remain tied to capabilities acquired during training or direct correspondence with annotated supports, motivating more flexible transfer paradigms.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

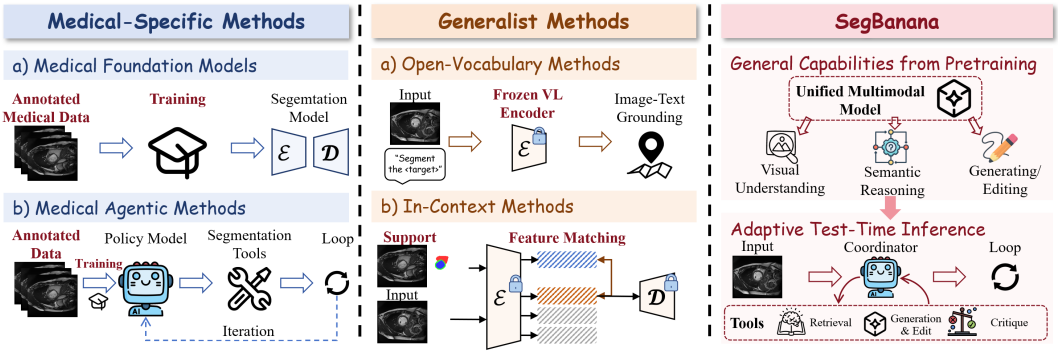


Figure 1: Comparison of generalization paradigms for medical image segmentation, including medical-specific methods, generalist methods, and our SegBanana.

Recent advances in large language models have shown that large-scale generative pretraining enables broad task transfer without task-specific updates (Brown et al., 2020; Chowdhery et al., 2023). Inspired by this, we ask: *can unified multimodal models (UMMs), which are similarly pretrained at scale with visual understanding, reasoning, and image generation capabilities, transfer to diverse medical segmentation tasks?* Unlike conventional segmentation models that rely on task-specific prediction spaces, UMMs define segmentation tasks through a shared generative interface conditioned on instructions and visual context, potentially allowing pretrained capabilities to be recomposed for new tasks at test time. Recent studies on natural images, including PixelArena (Liang et al., 2025) and VisionBanana (Gabeur et al., 2026), provide encouraging evidence that frontier UMMs already possess basic zero-shot segmentation capability and can be further extended to broader vision tasks through instruction tuning. However, how to effectively unlock this intrinsic capability for medical image segmentation remains unclear.

Our study reveals that UMMs already exhibit basic zero-shot medical segmentation capability, particularly on tasks with salient global shape or appearance cues. However, their performance degrades substantially on anatomically demanding tasks that require fine-grained discrimination of adjacent or nested structures, indicating that generic visual knowledge alone is insufficient when accurate segmentation depends on specialized anatomical and structural understanding. We further find that visual anatomical knowledge from in-context exemplars can compensate for insufficient task-specific priors, repeated sampling and diverse conditioning can expand the accessible prediction space, and targeted editing can effectively refine suboptimal predictions when sufficiently informative correction guidance is available.

Motivated by these observations, we propose SegBanana, a training-free medical segmentation agent that harnesses a frozen UMM through stateful, feedback-driven tool use. By casting medical image segmentation as structured visual generation, SegBanana uses a frozen UMM to generate and refine segmentation masks, augmented by Anatomy-Aware Knowledge Retrieval to supplement missing task-specific knowledge and Comparative Quality Critique to evaluate the quality of current masks and diagnose residual errors. A State-Aware Multimodal Controller reasons over the structured interaction trajectory and tool observations, and adaptively orchestrates retrieval, generation, editing, and critique to iteratively improve the segmentation result. Our contributions are three-fold:

- We systematically study the intrinsic medical segmentation capability of UMMs and identify knowledge augmentation, output exploration, and targeted editing as effective mechanisms for training-free transfer.
- We introduce SegBanana, a training-free medical segmentation agent that harnesses a frozen UMM through stateful, feedback-driven tool use, dynamically orchestrating knowledge retrieval, visual generation and editing, and comparative critique.
- Experiments on eight datasets show that SegBanana substantially improves frozen UMMs, achieves strong generalization across diverse clinical scenarios, and remains robust when only out-of-domain visual supports are available.

2 Related Work

Medical Image Segmentation Models. Existing medical segmentation models (Ma et al., 2024; Zhu et al., 2024; Zhao et al., 2025; Wang et al., 2025a; Magg et al., 2024) achieve broad coverage but still rely on task-specific training, which limits adaptation to unseen domains. Recent agentic medical segmentation frameworks (Jiang et al., 2026b; Huang et al., 2026) enable iterative segmentation, but depend on trained policy models and segmentation tools, limited by their training distributions. General open-vocabulary segmentation models (Barsellotti et al., 2025; Zeng et al., 2024; Carion et al., 2026) improve flexibility, but their performance is often constrained by limited medical knowledge and dependence on user-provided prompts. In-context segmentation methods (Cuttano et al., 2026; Zhang et al., 2023; 2024; Wang et al., 2023a;b) further adapt to new tasks through visual demonstrations, yet largely rely on feature correspondence and remain sensitive to the provided context. Instead, we explore an understanding-driven transfer paradigm that reasons over task semantics and visual context for training-free adaptation to unseen clinical scenarios.

Unified Multimodal Models. Unified multimodal models (UMMs) integrate visual understanding and generation within a framework (Team, 2024; Xie et al., 2025; Wang et al., 2024; Wu et al., 2025; Deng et al., 2025; Comanici et al., 2025; Lin et al., 2025). Chameleon (Team, 2024), Show-o (Xie et al., 2025), and Emu3 (Wang et al., 2024) unify multimodal modeling through shared token prediction, while Janus (Wu et al., 2025) decouples visual representations for understanding and generation within a unified backbone. Later models, including Nano Banana, extend these capabilities toward multimodal reasoning and instruction-guided image editing (Deng et al., 2025; Comanici et al., 2025; Lin et al., 2025). Building on pretrained UMMs, a growing line of work adapts them to general-purpose vision tasks through task-level supervision and post-training (Gabeur et al., 2026; Han et al., 2026). In contrast, we investigate whether the intrinsic visual capabilities of pretrained UMMs can be unlocked for medical image segmentation through agentic inference alone.

Agentic Frameworks for Image Generation and Editing. Agentic frameworks for image generation (Venkatesh et al., 2025; Nabati et al., 2024; Wang et al., 2025b; Jiang et al., 2026a; Shen et al., 2026) enhance generative models through multi-agent collaboration, preference-aware interaction, self-evaluation, and iterative tool use, enabling more diverse generation, improved user alignment, and finer-grained control over generated content. Agentic image editing frameworks (Zeng et al., 2026; Zhang et al., 2026; Zhao et al., 2026) further introduce iterative reasoning, collaborative planning and reflection, and reinforcement-learned coordination to improve edit quality, controllability, and alignment with complex instructions. These methods primarily use agentic inference to improve generation or editing within their original task settings. In contrast, we investigate whether agentic inference can transfer the capabilities of pretrained UMMs beyond generation and editing to training-free medical image segmentation.

3 Observation

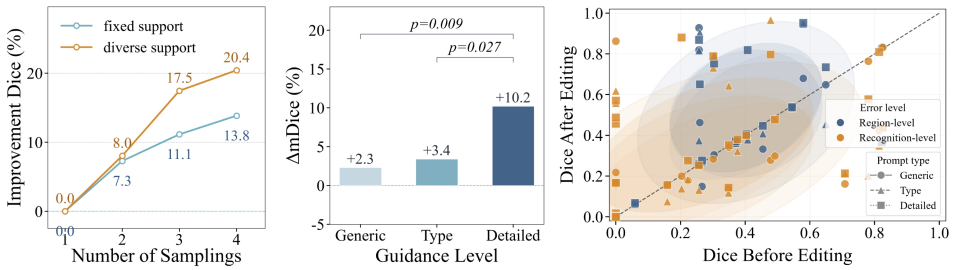
In this section, we characterize the intrinsic segmentation behavior of UMMs. Viewing segmentation as conditional generation, we study how external conditioning and inference-time sampling strategies shape the accessible output space. Moreover, leveraging the inherent editing capability of these models, we investigate whether existing predictions can be further revised and improved.

Generation Capability: *UMMs exhibit basic medical segmentation capability, but remain limited on tasks requiring specialized anatomical knowledge.* We evaluate UMMs on three representative tasks spanning different modalities, anatomical targets, scales, and spatial structures, as summarized in Tab. 1. For each task, the model receives only the input image and a task-specific instruction, e.g., "Segment the `<target>`", and directly produces a segmentation mask in a zero-shot manner. Then, the generated image is converted into a standardized discrete mask through pixel-wise color clustering. As shown in Tab. 1, UMMs perform substantially better on tasks with salient global shape or appearance cues, such as ISIC, than on ACDC and Drishti-GS, which require finer discrimination of adjacent or nested anatomical structures. These results suggest that UMMs can leverage generic visual cues for medical segmentation, but remain limited when accurate prediction requires specialized anatomical and structural knowledge.

Generation Conditioning: *Relevant visual supports effectively compensate for insufficient task-specific knowledge.* We provide image–mask exemplars as visual supports to supplement the UMMs

Table 1: Task characteristics and performance under text-only and support-augmented prompting, where Random samples a support and Retrieved selects a query-relevant one.

Dataset	Task Characteristics					mDice (%)		
	Modality	Target	Size	Structure	Main Challenge	Text-only	Random	Retrieved
ACDC (Bernard et al., 2018)	MRI	RV / MYO / LV	Small–Med.	Adjacent	Anatomical coupling	36.89	31.79	**45.30**
Drishti-GS (Sivaswamy et al., 2014)	Fundus	Disc / Cup	Small	Nested	Subtle boundaries	43.73	53.57	**57.49**
ISIC (Codella et al., 2019)	Dermoscopy	Lesion	Med.–Large	Irregular	Appearance variation	74.68	79.62	**80.65**



(a) (b) (c)

Figure 2: Inference-time exploration and refinement. (a) Best-of-N performance with increasing sampling. (b) Improvement under varying editing guidance. (c) Before–after Dice across error levels and prompt types, with shaded regions indicating their distributions.

with task-relevant anatomical knowledge. To examine how support selection affects segmentation, we compare randomly selected exemplars with query-relevant supports retrieved using DINO features (Siméoni et al., 2025). As shown in Tab. 1, query-relevant supports yield consistent improvements on tasks requiring specialized anatomical knowledge, whereas random supports provide less reliable gains and may introduce irrelevant context.

Generation Exploration: *Repeated sampling reveals diverse predictions, while varying the conditioning further expands the accessible output space.* We first examine whether stochastic generation can expose better segmentation predictions beyond a single output. On ACDC, we sample four outputs under the same query, instruction, and visual support, and evaluate oracle Best-of-N performance. As shown in Fig. 2(a), Best-of-N steadily improves as the sampling budget increases, indicating that even fixed conditioning yields substantial output diversity and that a single generation may overlook better predictions. We then vary the retrieved support while keeping the generation budget fixed. As shown in Fig. 2(a), sampling across different supports consistently outperforms repeated sampling with a fixed support, suggesting that conditioning diversity further broadens the accessible prediction space and exposes higher-quality candidate predictions.

Editing Behavior: *Detailed correction guidance substantially improves refinement, while editing is most reliable for region-level errors.* Beyond generating alternative predictions, UMMs can refine existing segmentations through editing. We study this capability on ACDC using the query image, a prediction, and a correction instruction. To examine the effect of instruction specificity, we compare three levels of guidance under the same image and mask: *generic editing*, which only requests correction; *type-guided editing*, which specifies the error type; and *detailed editing*, which localizes the error and describes the desired revision. As shown in Fig. 2(b), more specific guidance yields larger improvements, indicating that effective refinement benefits from precise and actionable instructions. We further analyze editing effectiveness across error types. Region-level errors preserve target identity and overall structure, whereas recognition-level errors involve incorrect identification or localization. As shown in Fig. 2(c), editing is more reliable for region-level errors, suggesting that it is better suited to refining plausible predictions than recovering fundamentally incorrect ones.

Further analyses of these behaviors are provided in Sec. B.5, including diversity-aware retrieval and the complementary effects of editing and resampling. Together, these observations motivate an agentic controller that can dynamically decide when to acquire knowledge, explore new masks, or refine an existing prediction according to the evolving segmentation state.

4 Method

To unlock the potential segmentation capability of pretrained UMMs, we cast medical image segmentation as structured visual generation and augment a frozen UMM with external anatomical

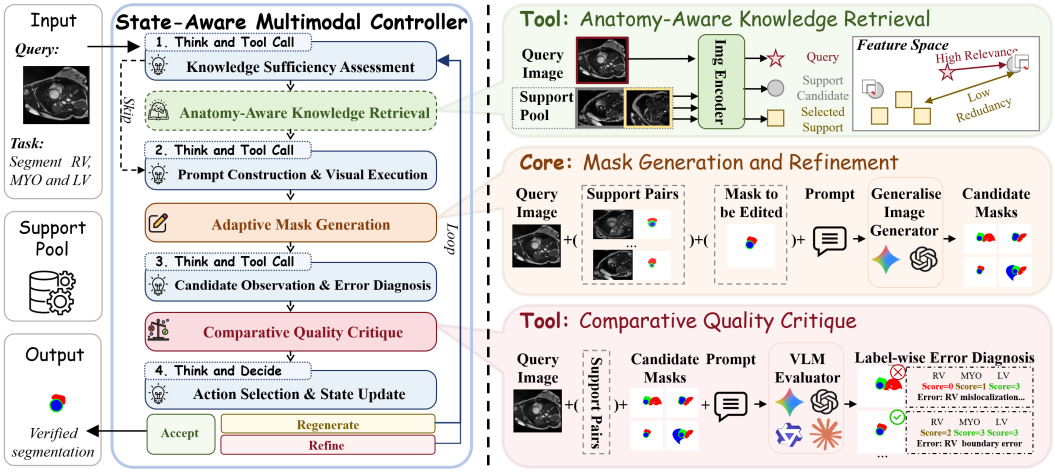


Figure 3: Overview of SegBanana. A State-Aware Multimodal Controller orchestrates Anatomy-Aware Knowledge Retrieval, UMM-based Mask Generation and Refinement, and Comparative Quality Critique, using state and feedback to adapt inference.

knowledge and comparative feedback under state-aware agentic control, enabling iterative exploration and refinement of segmentation predictions.

Method Overview. We present an overview of SegBanana in Fig. 3. Given a query image, a segmentation instruction, and an external pool of annotated visual supports, SegBanana performs segmentation through closed-loop visual generation and refinement centered on a frozen UMM. At each iteration, a State-Aware Multimodal Controller reasons over the current state and observations to select the next tool action. When additional task-specific anatomical knowledge is required, it invokes Anatomy-Aware Knowledge Retrieval to acquire relevant visual support. The controller then synthesizes an action-specific instruction and invokes the frozen UMM to either explore a new segmentation mask or edit the retained prediction. The resulting candidates are then examined by Comparative Quality Critique, whose preference and residual-error diagnosis are returned as feedback. This feedback updates the structured state and closes the loop, allowing the controller to reassess the current prediction and adapt its next action until Accept is selected or the inference budget is exhausted.

4.1 Capability-Enhancing Segmentation Toolset

Anatomy-Aware Knowledge Retrieval. Given a query image I, task instruction T, and support pool $\mathcal{D}_s = \{(I_i, M_i)\}_{i=1}^{N}$, we adopt maximal marginal relevance (MMR) (Carbonell & Goldstein, 1998) to balance query relevance and support redundancy:

$$i_k = \arg\max_{i \notin \mathcal{S}_{k-1}} \left[\lambda \, s_q(i) - (1-\lambda) \max_{j \in \mathcal{S}_{k-1}} s(i,j) \right], \tag{1}$$

where $s_q(i)$ measures query–support similarity and $s(i,j)$ measures similarity between support examples. After K selections, the retriever returns

$$\mathcal{S}_t = \mathcal{R}(I, T, \mathcal{D}_s) = \{(I_{i_k}, M_{i_k})\}_{k=1}^{K}, \tag{2}$$

which provides task-relevant anatomical, semantic, and appearance cues for subsequent generation.

Mask Generation and Refinement. Given a query image I and task instruction T, the controller invokes the frozen UMM through two visual actions:

$$M_t = \begin{cases} G(I, T, \mathcal{K}_t), & \textsc{Generate}, \\ G(I, B_t, E_t), & \textsc{Edit}, \end{cases} \tag{3}$$

where $\mathcal{K}_t$ denotes the context, which may include retrieved image–mask supports, B_t is the retained best prediction, and E_t is a targeted correction instruction synthesized from the current state. The two modes enable exploration of alternative predictions and local refinement of promising results.

Comparative Quality Critique. We employ a frozen multimodal VLM as a feedback tool that comparatively evaluates competing predictions and diagnoses residual errors. Let $\mathcal{M}_t = \{M_t^{(1)}, \ldots, M_t^{(n)}\}$ denote the newly produced candidates. If no retained prediction is available, the candidates are compared sequentially to determine a provisional champion. Otherwise, the retained best prediction B_t is introduced as a baseline and challenged by the new candidates:

$$(\hat{M}_t, O_t^{\text{crit}}) = \text{Tournament}_{\mathcal{V}}\left(I,\, \mathcal{M}_t,\, B_t\right), \tag{4}$$

where $\hat{M}_t$ denotes the tournament winner and O_t^{crit} contains the corresponding visual evaluation and residual-error diagnosis. The critic compares predictions based on target identity, location, shape, topology, and boundary alignment. For multi-label tasks, different labels may be retained from different candidates and composed into the current prediction. The resulting preference and error diagnosis are returned as structured feedback, forming the observation that drives the controller's next action.

4.2 State-Aware Multimodal Controller

SegBanana employs a frozen multimodal VLM as a ReAct-style controller that reasons over the query, task instruction, stage-relevant observations, and compact interaction trajectory to determine the next tool action and refinement strategy.

Structured State. Rather than replaying the full interaction trajectory, SegBanana maintains a compact structured state

$$\mathcal{Z}_t = \{\mathcal{C}_t, \mathcal{H}_t\}, \tag{5}$$

comprising a *Current State* $\mathcal{C}_t$ and a *Historical State* $\mathcal{H}_t$. The Current State summarizes the observations required for the next action decision, including the latest prediction, visual evaluation, residual-error diagnosis. The Historical State summarizes cross-iteration information:

$$\mathcal{H}_t = \left\{B_t,\, \mathcal{E}_t,\, \mathcal{H}_t^{\text{edit}},\, \mathcal{H}_t^{\text{ret}}\right\}, \tag{6}$$

where B_t denotes the retained best prediction, $\mathcal{E}_t$ summarizes accumulated error observations, $\mathcal{H}_t^{\text{edit}}$ tracks previous editing attempts and outcomes, and $\mathcal{H}_t^{\text{ret}}$ tracks explored supports and their observed utility. This compact state serves as the agent's working memory, preserving cross-iteration experience while exposing only stage-relevant information to each module, enabling the controller to track unresolved errors and avoid ineffective actions without replaying the full interaction trajectory.

State-Aware Tool Use. At each iteration, the controller first determines whether additional task-specific anatomical knowledge is needed from the query, task specification, and compact policy context. If needed, Anatomy-Aware Knowledge Retrieval is invoked; otherwise, the existing visual context is reused. The controller then synthesizes an executable instruction for the selected action. For generation, it uses the query and, when available, a selected visual support; for editing, it reasons over the retained best prediction, diagnosed errors, and relevant interaction trajectory to formulate a targeted correction. The image model receives only the synthesized instruction and action-specific visual inputs. Each new candidate is evaluated by Comparative Quality Critique, whose preference and residual-error diagnosis are incorporated into the structured state. The retained best prediction is updated as

$$B_{t+1} = \begin{cases} \hat{M}_t, & \hat{M}_t \succ B_t, \\ B_t, & \text{otherwise}, \end{cases} \tag{7}$$

where $\succ$ denotes the preference from comparative critique. Without a previous best, the tournament winner initializes B_t. This best prevents weaker generations or unsuccessful edits from overwriting stronger predictions. Stage-specific prompts are provided in Sec. B.4.

Adaptive Action Selection. Based on the current visual diagnosis, prior policy context, and interaction trajectory, the controller operates over an action space $\mathcal{A} = \{\text{GENERATE}, \text{EDIT}, \text{ACCEPT}\}$. GENERATE explores a new prediction when the retained result remains globally unreliable; EDIT refines the current best when it provides a reliable basis for local correction; and ACCEPT terminates inference once the result is satisfactory. Otherwise, the updated state is carried to the next iteration, where knowledge needs and refinement strategy are reassessed. Inference also stops when the iteration budget is exhausted, returning the best retained prediction.

Table 2: Comparison of SegBanana against medical-specific and generalist baselines across eight medical segmentation datasets in terms of mDice (%). ♦/▲denote zero-/few-shot settings. Gray denotes medical-specific results lower than SegBanana. **Bold**/underline indicate the best/second-best results within each method category.

Method	TNBC	RAVIR	Drishti-GS	ISIC	BUS-UCLM	Kvasir	ACDC	BraTS	Avg
Medical-Specific Methods									
UniverSeg▲	24.48	**28.45**	40.68	39.64	13.38	19.91	21.48	15.48	25.44
BiomedParse♦	11.81	5.67	**78.19**	87.59	**63.81**	88.95	**89.03**	**75.11**	**62.52**
MedSAM3♦	**68.21**	4.21	34.13	**89.80**	57.68	**91.17**	16.74	55.42	52.17
IBISAgent♦	1.00	1.65	4.14	87.47	57.06	78.86	14.87	60.14	38.15
MedSAMAgent♦	5.78	0.30	66.41	83.47	49.16	83.92	80.50	21.38	48.87
Generalist Methods									
Talk2DINO♦	21.73	22.23	2.16	26.69	12.14	23.35	4.74	6.74	14.97
SED♦	22.05	22.71	1.76	30.59	21.45	29.52	3.50	14.78	18.30
MaskCLIP++♦	23.25	22.21	1.50	28.61	12.57	23.15	1.53	5.09	14.74
SAM3♦	0.00	0.00	1.79	22.98	5.85	0.00	56.65	0.00	10.91
Painter▲	22.94	6.55	1.72	27.65	8.05	27.60	1.36	15.55	13.93
SegGPT▲	53.33	64.19	76.82	51.81	32.08	65.57	45.20	29.08	52.26
Matcher▲	25.17	21.24	35.61	55.09	38.38	53.67	20.48	27.94	34.70
PerSAM▲	5.25	19.41	33.37	28.53	32.46	44.99	18.19	14.68	24.61
GF-SAM▲	48.98	37.91	24.48	63.68	42.07	53.00	13.16	17.20	37.56
INSID3▲	31.16	33.63	6.25	63.87	23.72	74.49	4.78	21.72	32.45
SegBanana♦▲	**74.75**	**80.93**	**87.09**	**82.06**	**78.07**	**89.53**	**61.68**	**65.52**	**77.45**

5 EXPERIMENTS

In this section, we evaluate SegBanana on eight medical segmentation datasets and compare it with a broad range of baselines under a training-free setting. The datasets span diverse imaging modalities and segmentation targets: TNBC (Naylor et al., 2018) for nuclei in histopathology, RAVIR (Hatamizadeh et al., 2022) for retinal vessels in fundus imaging, Drishti-GS for optic disc and cup in fundus imaging, ISIC for skin lesions in dermoscopy, BUS-UCLM (Vallez et al., 2025) for breast lesions in ultrasound, Kvasir (Jha et al., 2019) for gastrointestinal polyps in endoscopy, ACDC for cardiac structures in MRI, and BraTS (Menze et al., 2014; Bakas et al., 2017; Baid et al., 2021) for brain tumors in multimodal MRI, with details in Sec. B.1. We follow standard splits and report mean Dice (mDice) (Dice, 1945) as the primary metric. We use Nano Banana 2 for mask generation and editing, Gemini-3.5 as the multimodal controller, and DINOv3 ViT-L/16 with MMR ($\lambda = 0.8$) for support retrieval. Each Generate action produces two candidates, each Edit action refines the current best prediction, and inference runs for up to three iterations with a compact structured state.

We compare against both medical-specific and generalist methods. Medical-specific baselines include UniverSeg (Butoi et al., 2023), BiomedParse (Zhao et al., 2025), MedSAM3 (Liu et al., 2025), IBISAgent (Jiang et al., 2026b), and MedSAMAgent (Liu et al., 2026). Generalist baselines include open-vocabulary and in-context methods: Talk2DINO (Barsellotti et al., 2025), SED (Xie et al., 2024), MaskCLIP++ (Zeng et al., 2024), SAM3 (Carion et al., 2026), Painter (Wang et al., 2023a), SegGPT (Wang et al., 2023b), Matcher (Liu et al., 2023), PerSAM (Zhang et al., 2023), GF-SAM (Zhang et al., 2024), and INSID3 (Cuttano et al., 2026). We exclude SAM/SAM2-based methods requiring point or box prompts, as they introduce extra supervision unavailable in our setting.

5.1 MAIN RESULTS

SegBanana achieves strong and balanced performance across diverse medical segmentation tasks. As shown in Tab. 2, SegBanana achieves the highest average mDice of 77.45%, outperforming representative medical-specific and generalist baselines, BiomedParse and SegGPT, by 14.93 and 25.19 points, respectively. It performs best on four datasets, including TNBC, RAVIR, Drishti-GS, and BUS-UCLM, and remains competitive with medical-specific methods on the other four. Notably, several medical-specific methods outperforming SegBanana on these datasets have dataset-level exposure to the corresponding benchmarks during training (Sec. B.2), underscoring the competitiveness of SegBanana in a training-free setting. Across method families, medical-specific

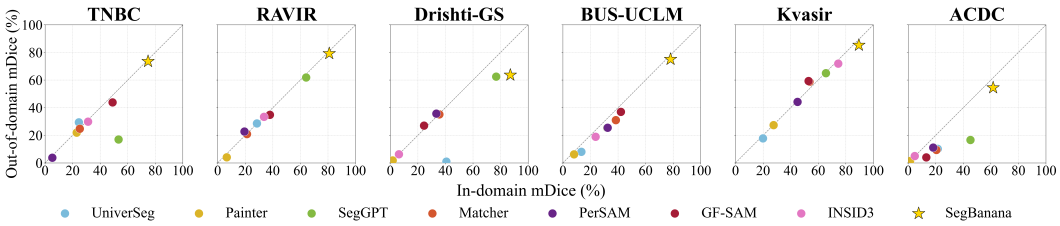


Figure 4: In-domain vs. out-of-domain mDice (%) across datasets. Points closer to the diagonal indicate greater robustness to support-domain shift.

Table 3: Ablation of framework components and multimodal controller backbones across eight datasets. Results are reported in mDice (%).

Framework	Controller	TNBC	RAVIR	Drishti-GS	ISIC	BUS-UCLM	Kvasir	ACDC	BraTS	Avg.
SegBanana	Gemini-3.5	74.75	80.93	87.09	82.06	78.07	89.53	61.68	65.52	77.45
w/o Retrieval	Gemini-3.5	76.26	81.22	60.55	82.78	77.92	88.19	51.56	61.67	72.52
w/o Critic	Gemini-3.5	72.08	78.05	63.88	81.16	75.07	86.02	56.68	65.66	72.33
w/o State Context	Gemini-3.5	73.86	77.30	65.07	81.85	76.18	87.74	56.36	63.43	72.72
SegBanana	Qwen-3.8	70.06	41.94	83.28	79.03	69.37	82.31	51.31	54.32	66.45
SegBanana	GPT-5.6	62.38	54.36	88.15	82.78	77.54	85.39	55.39	61.88	70.98
Nano Banana 2	N/A	67.04	34.44	43.73	74.68	69.30	85.06	36.89	50.38	57.69

models perform strongly within their training coverage but degrade beyond it, generalist methods may lack specialized anatomical knowledge and suffer from foreground–background ambiguity and coarse patch-level matching, limiting fine-structure segmentation. In contrast, SegBanana remains consistently strong across all datasets.

SegBanana remains robust when in-domain supports are unavailable. New clinical scenarios may lack matched annotated supports, while related external datasets with similar targets remain available. We therefore evaluate six query datasets using suitable public out-of-domain supports and compare representative few-shot methods under the same setting, with dataset pairings provided in Sec. B.3. As shown in Fig. 4, points near the diagonal indicate robustness to support-domain shift. SegBanana remains close to the diagonal on most datasets, whereas several baselines degrade substantially. For example, SegGPT drops from 53% to 17% on TNBC and from 45% to 16% on ACDC. These results highlight the potential of SegBanana to reduce the annotation burden for deploying medical segmentation in new clinical practice.

5.2 Ablation

SegBanana consistently benefits from its key components and remains robust to the choice of multimodal controller. We assess the contribution of each framework component and the robustness of SegBanana to multimodal controllers. As shown in Tab. 3, *w/o retrieval*, *w/o critic*, and *w/o state context* reduce average mDice from 77.45% to 72.52%, 72.33%, and 72.72%, respectively. Here, *w/o retrieval* forces all cases to skip support retrieval; *w/o critic* integrates comparative evaluation into the controller rather than using a separate critic tool; and *w/o state context* replaces the compact structured state with the full interaction context. Retrieval is particularly important on knowledge-intensive datasets such as Drishti-GS and ACDC, while the critic and state ablations demonstrate the benefits of dedicated comparative evaluation and compact history-aware reasoning. SegBanana also remains effective with Qwen and GPT controllers despite variation across backbones. Finally, direct generation with Nano Banana 2 achieves only 57.69% average mDice, highlighting the substantial gain brought by agentic inference over the frozen generator alone.

5.3 Discussion

Closed-loop inference progressively improves segmentation performance, but Comparative Quality Critique remains limited in consistently identifying the best generated candidate. Fig. 5 shows that closed-loop inference steadily expands the pool of strong predictions, as reflected by the consistently improving oracle performance. While SegBanana allows up to three iterations,

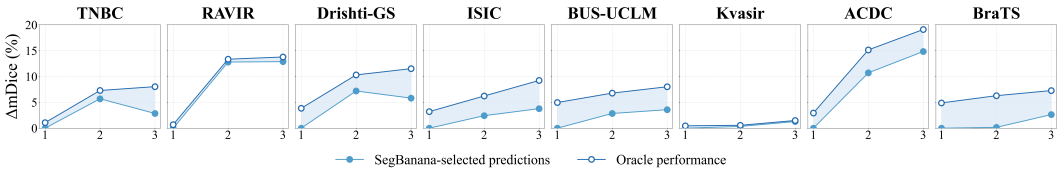


Figure 5: Iterative segmentation performance across inference steps, with the ΔmDice of critique-selected predictions and the oracle performance among all masks generated up to each iteration.

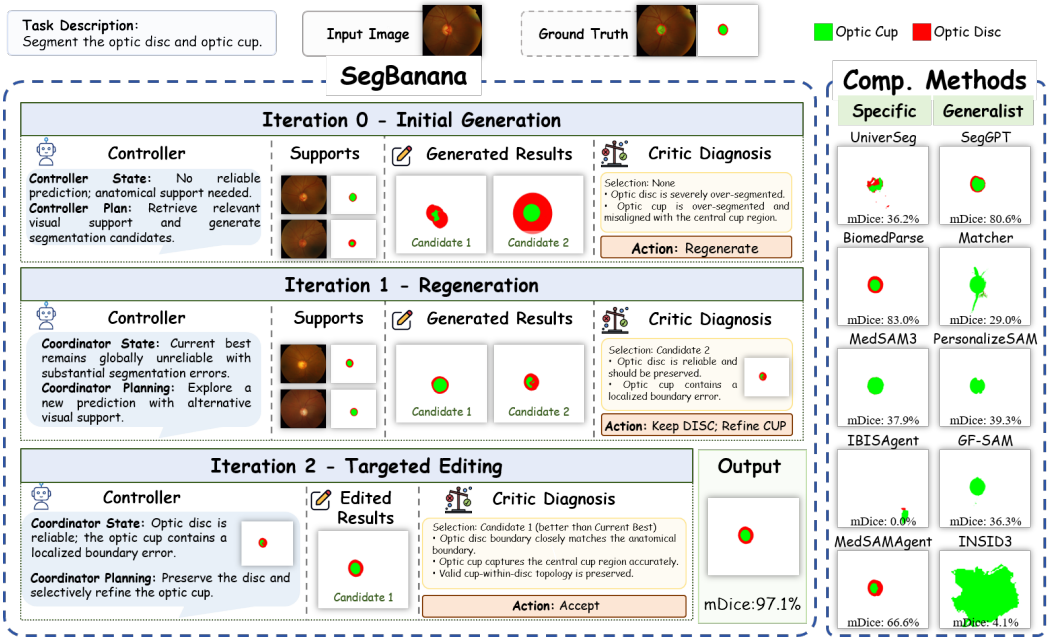


Figure 6: Case study of SegBanana, illustrating its iterative segmentation process and comparison with the top five methods from medical-specific and generalist baselines.

its actual inference depth and action usage are adaptively determined by the evolving prediction state, with detailed efficiency and stopping statistics provided in Sec. B.6. However, the prediction selected by Comparative Quality Critique remains below the oracle across iterations, indicating that better candidates are often generated but not reliably identified. This gap highlights candidate selection as a key remaining bottleneck and suggests that improving training-free comparative evaluation could further unlock the gains already available from iterative generation and refinement.

Qualitative Study. Fig. 6 illustrates how SegBanana benefits from adaptive, state-aware inference. When the initial prediction suffers severe over-segmentation, the critic triggers regeneration rather than local editing. Once a stronger candidate is obtained, the controller preserves the reliable optic disc and selectively refines the remaining cup error, yielding a final mDice of 97.1%. This trajectory illustrates how the controller adapts its action policy as the state evolves, switching from global exploration to local refinement once a reliable hypothesis emerges. Additional case studies across diverse datasets are provided in Sec. B.7.

6 Conclusion

In this paper, we present SegBanana, an agentic visual generation framework that enables a frozen UMM to perform medical image segmentation without task-specific post-training. SegBanana harnesses intrinsic generation and editing capabilities through retrieval, refinement, and verification at inference time. Across diverse datasets, it outperforms medical-specific and generalist baselines and remains robust without in-domain supports, demonstrating effective training-free transfer. A key **limitation** is the imperfect reliability of comparative quality critique, which prevents SegBanana from fully exploiting the UMM's segmentation potential. **Future work** will pursue stronger verification and more autonomous agentic control, including a dedicated policy model for reasoning over intermediate outcomes and orchestrating visual actions.

AI USE STATEMENT

For manuscript preparation, we use OpenAI GPT-family models only for language polishing and grammar correction. Generative and multimodal models used as experimental components of SegBanana are explicitly described in the paper. AI tools were not used for literature review or idea formation.

ETHICS STATEMENT

This work uses only publicly available medical image segmentation datasets, and no new patient data were collected, no human subjects were recruited, and no patient interaction occurred. We do not attempt to identify or re-identify individuals from the medical images. SegBanana is intended solely for research purposes and is not a medical device. Although the framework demonstrates strong transfer across diverse datasets, segmentation errors may still arise under challenging anatomy, distribution shifts, or imperfect candidate evaluation. We therefore report methodological limitations and do not claim clinical readiness.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed implementation settings in Sec. 5 and the supplementary material, including model configurations, stage-specific prompts and input–output formats, support retrieval, structured state construction, agentic rollouts, and hyperparameters. We specify the UMM and multimodal controller used in each experiment, the DINOv3-based retrieval and MMR settings, the number of candidates generated per action, and the maximum number of inference iterations. Upon publication, we will release the code, prompt templates, configuration files, dataset splits, support-pool construction, and preprocessing pipeline. Because SegBanana relies on externally hosted UMM APIs for generation, editing, and multimodal reasoning, exact outputs may vary slightly due to stochastic inference or future API/model updates. To mitigate this source of variation, we will document the exact model versions and inference settings used for all reported results.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):170117, 2017.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22025–22035. IEEE, 2025.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 21381–21394. IEEE, 2023.

Jaime G Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, volume 98, pp. 290941–291025, 1998.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, et al. Vision as unified multimodal generation. *arXiv preprint arXiv:2607.06560*, 2026.

Ali Hatamizadeh, Hamid Hosseini, Niraj Patel, Jinseo Choi, Cameron C Pole, Cory M Hoeferlin, Steven D Schwartz, and Demetri Terzopoulos. Ravir: A dataset and methodology for the semantic segmentation and quantitative analysis of retinal arteries and veins in infrared reflectance imaging. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3272–3283, 2022.

Ziyan Huang, Haoyu Wang, Jin Ye, Yuanfeng Ji, Xiaowei Hu, Lihao Liu, Zhikai Yang, Wei Li, Ming Hu, Yanzhou Su, et al. Medsegagent: A universal and scalable multi-agent system for instructive medical image segmentation. *IEEE journal of biomedical and health informatics*, 2026.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. Genagent: Scaling text-to-image generation via agentic multimodal reasoning. *arXiv preprint arXiv:2601.18543*, 2026a.

Yankai Jiang, Qiaoru Li, Binlu Xu, Haoran Sun, Chao Ding, Junting Dong, Yuxiang Cai, Xuhong Zhang, and Jianwei Yin. Ibisagent: Reinforcing pixel-level visual reasoning in mllms for universal biomedical object referring and segmentation. *arXiv preprint arXiv:2601.03054*, 2026b.

Feng Liang, Sizhe Cheng, Chenqi Yi, and Yong Wang. Pixelarena: A benchmark for pixel-precision visual intelligence. *arXiv preprint arXiv:2512.16303*, 2025.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Anglin Liu, Rundong Xue, Xu R Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. Medsam3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025.

Shengyuan Liu, Liuxin Bao, Qi Yang, Wanting Geng, Boyun Zheng, Chenxin Li, Wenting Chen, Houwen Peng, and Yixuan Yuan. Medsam-agent: Empowering interactive medical image segmentation with multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2602.03320*, 2026.

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

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Caroline Magg, Hoel Kervadec, and Clara I. Sánchez. Zero-shot capability of sam-family models for bone segmentation in ct scans, 2024. URL `https://arxiv.org/abs/2411.08629`.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Ofir Nabati, Guy Tennenholtz, ChihWei Hsu, Moonkyung Ryu, Deepak Ramachandran, Yinlam Chow, Xiang Li, and Craig Boutilier. Preference adaptive and sequential text-to-image generation. *arXiv preprint arXiv:2412.10419*, 2024.

Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38 (2):448–459, 2018.

Ziwei Niu, Shuyi Ouyang, Shiao Xie, Yen-wei Chen, and Lanfen Lin. A survey on domain generalization for medical image analysis. *arXiv preprint arXiv:2402.05035*, 2024.

Shaocheng Shen, Jianfeng Liang, Chunlei Cai, Cong Geng, Huiyu Duan, Xiaoyun Zhang, Qiang Hu, and Guangtao Zhai. Agentic retoucher for text-to-image generation. *arXiv preprint arXiv:2601.02046*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pp. 53–56. IEEE, 2014.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Noelia Vallez, Gloria Bueno, Oscar Deniz, Miguel Angel Rienda, and Carlos Pastor. Bus-uclm: Breast ultrasound lesion segmentation dataset. *Scientific Data*, 12(1):242, 2025.

Kavana Venkatesh, Connor Dunlop, and Pinar Yanardag. Crea: A collaborative multi-agent framework for creative content generation with diffusion models. *arXiv e-prints*, pp. arXiv–2504, 2025.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 2025a.

Kaishen Wang, Ruibo Chen, Tong Zheng, and Heng Huang. Imagent: A unified multimodal agent framework for test-time scalable image generation. *arXiv preprint arXiv:2511.11483*, 2025b.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839. IEEE, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3426–3436. IEEE, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Quan-Sheng Zeng, Yunheng Li, Daquan Zhou, Guanbin Li, Qibin Hou, and Ming-Ming Cheng. Maskclip++: A mask-based clip fine-tuning framework for open-vocabulary image segmentation. 2024.

Ziyun Zeng, Hang Hua, and Jiebo Luo. Mira: Multimodal iterative reasoning agent for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9563–9573, 2026.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. 2024.

R Zhang, Z Jiang, Z Guo, S Yan, J Pan, X Ma, H Dong, P Gao, and H Li. Personalize segment anything model with one shot. arxiv 2023. *arXiv preprint arXiv:2305.03048*, 2023.

Yuxuan Zhang, Shijia Huang, and Liwei Wang. Intentedit: Multi-agent reasoning for intent-driven complex image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8776–8785, 2026.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1): 166–176, 2025.

Yiran Zhao, Yaoqi Ye, Xiang Liu, Michael Qizhe Shieh, and Trung Bui. Imageedit-r1: Boosting multi-agent image editing via reinforcement learning. *arXiv preprint arXiv:2603.08059*, 2026.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

Table 4: Characteristics of the evaluation datasets in terms of imaging modality, segmentation target, spatial structure, and primary segmentation challenge.

Dataset	Modality	Target	Structure	Primary Challenge
ACDC	MRI	LV / RV / MYO	Adjacent	Anatomical coupling
Drishti-GS	Fundus	Optic disc / cup	Nested	Subtle boundary discrimination
ISIC	Dermoscopy	Skin lesions	Irregular	Appearance variation
TNBC	Histopathology	Cell nuclei	Irregular, Adjacent	Dense clustering & appearance variation
RAVIR	Fundus	Retinal vessels	Irregular	Subtle boundaries & topology preservation
BUS-UCLM	Ultrasound	Breast lesions	Irregular	Appearance variation & weak boundaries
Kvasir	Endoscopy	Colorectal polyp	Irregular	Appearance variation & specular reflection
BraTS	MRI	Brain tumor	Nested	Anatomical coupling & appearance variation

A Limitations and Future Work

SegBanana currently has two main limitations. First, the framework operates on 2D images and cannot directly model volumetric 3D context, requiring 3D scans to be processed slice by slice. Second, although Comparative Quality Critique enables training-free candidate selection and refinement, its reliability remains imperfect, preventing the framework from fully exploiting the segmentation capability of pretrained UMMs. Future work could incorporate targeted supervision to strengthen both the controller and the UMM on medical images, while preserving their pretrained visual understanding, reasoning, and generation abilities. Such adaptation may further improve medical segmentation performance without sacrificing generality, moving toward a more flexible and broadly transferable segmentation paradigm.

B Appendix

B.1 Dataset Details and Coverage

We evaluate SegBanana on eight medical image segmentation datasets covering diverse imaging modalities, anatomical targets, spatial scales, and structural characteristics. As summarized in Tab. 4, these datasets span MRI, fundus imaging, dermoscopy, histopathology, ultrasound, and endoscopy, with targets ranging from small cellular and vascular structures to large lesions and multi-region anatomical structures. They also exhibit substantially different spatial organizations, including adjacent, nested, and irregular structures, resulting in distinct segmentation challenges such as subtle boundaries, anatomical coupling, dense object clustering, topology preservation, and large appearance variation. Such diversity allows us to evaluate not only cross-modality generalization, but also the ability to transfer across segmentation tasks with different requirements for appearance recognition, boundary discrimination, and anatomical reasoning.

B.2 Supervision Profile of Compared Methods

Medical-specific baselines differ substantially in the amount and diversity of medical data observed during model development. To better contextualize their performance, we summarize the publicly reported medical data exposure of each method in Tab. 5. We consider a dataset as "seen" whenever it is explicitly used at any stage of model development, including pretraining, supervised fine-tuning (SFT), or reinforcement learning (RL), since exposure at any of these stages may influence downstream segmentation behavior. Dataset names highlighted in red indicate benchmark-level overlap with our evaluation datasets. This notation indicates exposure to the same benchmark and does not necessarily imply identical training/test samples or splits.

Overall, the compared medical-specific methods exhibit markedly different training exposure. Several methods have observed datasets from the same modalities, anatomical regions, or even the same benchmarks used in our evaluation. Such prior exposure provides useful context for interpreting dataset-specific performance and out-of-domain generalization. For methods whose complete foundation-model pretraining corpora are not publicly disclosed, we report only explicitly documented medical datasets or modality-level exposure and make no assumptions about undisclosed data.

Table 5: Reported medical data exposure of the compared medical-specific methods. Datasets used during any model-development stage, including pretraining, fine-tuning/SFT, and RL, are counted as exposure. Gray rows denote modalities included in our evaluation benchmark, while red dataset names indicate benchmark-level overlap with our evaluation datasets. ‡ indicates that exposure to the modality is explicitly reported, but the complete dataset-level training list is not publicly disclosed.

Modality	UniverSeg	BiomedParse	MedSAM3	IBISAgent	MedSAM-Agent
MRI	BraTS, BrainDevelopment, ISLES, LGG, MCIC, OASIS, PPMI, WMH, I2CVB, NCI-ISBI, PROMISE12, FeTA, SpineWeb, ACDC, CDEMRIS	BraTS2023, PROMISE12, LGG, ACDC, M&Ms	MRI corpus‡	ACDC, LGG, M&Ms, MSD Brain Tumor, MSD Heart, MSD Hippocampus, MSD Prostate	ACDC, LGG, AMOS-MRI
CT	AbdomenCT-1K, BTCV, KiTS, LiTS, LUNA, SegTHOR, WORD	AbdomenCT-1K, BTCV, KiTS23, TotalSegmentator, COVID-19 CT, LIDC-IDRI, FUMPE	CT corpus‡	COVID-19 CT, KiTS23, LIDC-IDRI, MSD Liver, MSD Spleen, MSD Pancreas, MSD Colon, MSD Lung, MSD Hepatic Vessel	FLARE22, KiTS, LIDC-IDRI, BTCV, AMOS-CT, WORD
MRI & CT	AMOS, CHAOS, MSD	AMOS22, MSD	–	AMOS22, MSD task groups	–
Fundus / Retinal	DRIVE, STARE, IDRID, e-Ophtha	DRIVE, REFUGE, G1020	RIM-ONE	G1020, REFUGE	REFUGE
X-ray	CheXplanation, PanDental	COVID-QU-Ex, QaTa-COV19, SIIM-ACR Pneumothorax, Chest Xray Masks and Labels, CDD-CESM, Radiography	X-ray corpus‡	COVID-QU-Ex, QaTa-COV19, Radiography, CDD-CESM, SIIM-ACR Pneumothorax, CXR Masks and Labels	CXRMask, Radiography series, CDD-CESM
Ultrasound	CAMUS, HMC-QU, BUS, TUCC	CAMUS, BUSI, US Simulation & Segmentation, BreastUS, LiverUS, FH-PS-AOP	BUSI; broader US corpus‡	CAMUS, BreastUS, LiverUS, FH-PS-AOP	BreastUS, LiverUS, FH-PS-AOP
Histopathology / Microscopy	BBBC003, WBC, CoNSeP	PanNuke, GlaS	Histopathology / microscopy corpus‡	GlaS, PanNuke	–
Dermoscopy	–	ISIC 2018, UWaterlooSkinCancer	ISIC 2018; broader corpus‡	ISIC, UWaterlooSkinCancer	–
Endoscopy	–	PolypGen, NeoPolyp	Kvasir-SEG; broader corpus‡	NeoPolyp, PolypGen	NeoPolyp, PolypGen
OCT / OCTA	OCTA500, ROSE	OCT-CME	OCT corpus‡	OCT-CME	–

Our evaluation datasets. MRI: ACDC and BraTS; Fundus: Drishti-GS and RAVIR; Ultrasound: BUS-UCLM; Histopathology: TNBC; Dermoscopy: ISIC; Endoscopy: Kvasir. For MedSAM3, the public release reports broader multimodal medical training than the individually named datasets; therefore, modality-level exposure is reported where the complete dataset list is unavailable.

B.3 Out-of-Domain Support Construction

To evaluate SegBanana when matched in-domain annotations are unavailable, we construct an out-of-domain support setting in which each query dataset is paired with an external annotated dataset containing the same or closely corresponding segmentation target. Importantly, the external supports are drawn from a different dataset rather than from the query distribution, introducing shifts in acquisition conditions, appearance, population, and annotation characteristics while preserving the

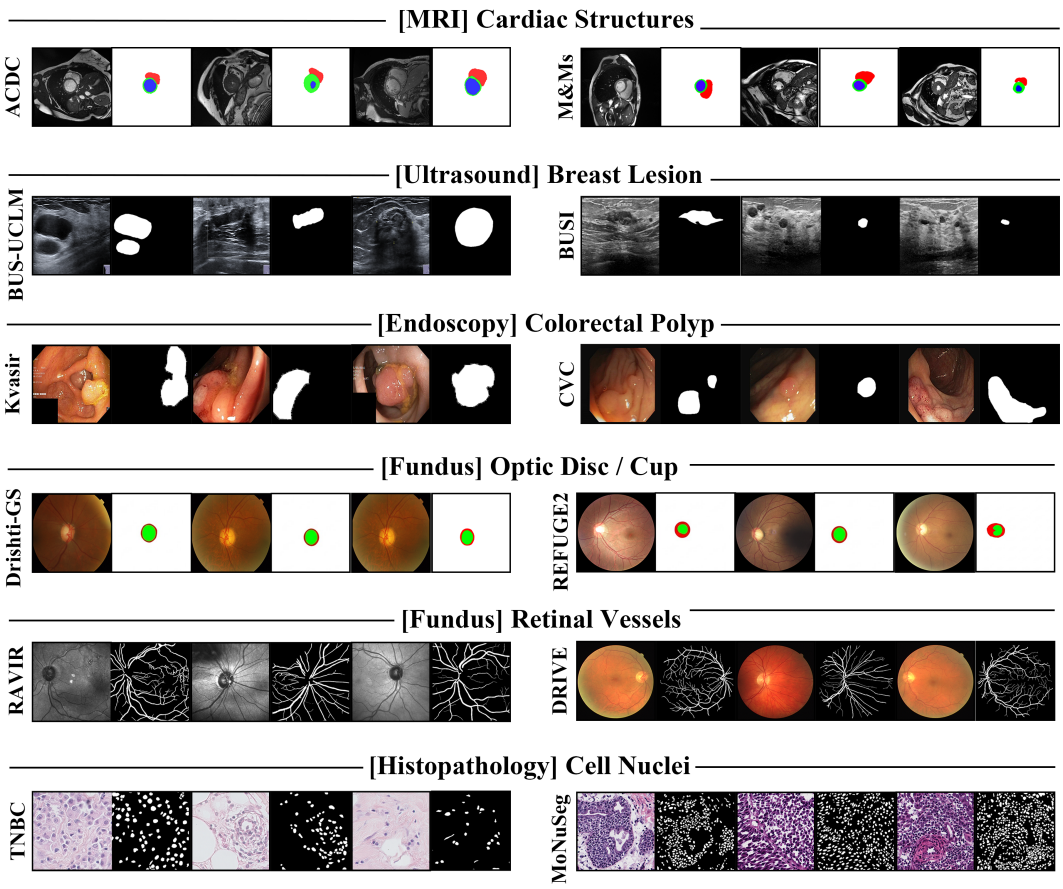


Figure 7: Out-of-domain support construction across six medical segmentation tasks. For each task, the query dataset (left) is paired with an external annotated dataset (right) sharing the same or closely corresponding segmentation target, while differing in acquisition conditions, appearance, population, or annotation characteristics. Example image–mask pairs illustrate the resulting domain shift while preserving task-relevant anatomical semantics.

task-relevant anatomical semantics. The dataset pairings are visualized in Fig. 7 to illustrate the corresponding domain gaps.

B.4 Prompt and Inference Details

SegBanana decomposes inference into several stage-specific prompting modules, including knowledge-sufficiency assessment, generation and editing instruction synthesis, visual execution, comparative critique, and action selection. Generation and editing each contain two stages: the controller first synthesizes an executable instruction from the relevant interaction state, and the frozen UMM then executes this instruction using only the visual inputs required by the selected action. Different stages therefore receive different projections of the structured state rather than the complete interaction trajectory. Controller-side prompts use fixed JSON output formats for deterministic parsing, while generation and editing execution directly output RGB segmentation masks.

Task Description. Each dataset is associated with a concise segmentation instruction defining the target structures, as summarized in Tab. 6. These task descriptions are further instantiated with dataset-specific label definitions and RGB conventions when constructing the corresponding prompts.

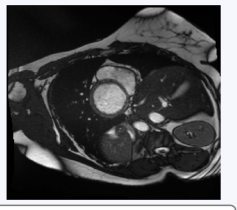
Knowledge Sufficiency Assessment. Before each generation step, the controller determines whether the available context contains sufficient task-specific anatomical knowledge. As illustrated

Table 6: Dataset-specific task descriptions used in prompts.

Dataset	Task Description
ACDC	Segment the right ventricular cavity, myocardium, and left ventricular cavity.
BUS-UCLM	Segment the breast lesion or mass.
Kvasir	Segment the gastrointestinal polyp.
RAVIR	Segment the retinal vessels.
BraTS	Segment the brain tumor.
TNBC	Segment the cell nuclei.
Drishti-GS	Segment the optic disc and optic cup.
ISIC	Segment the skin lesion.

Knowledge Sufficiency Assessment Prompt

<STATIC INSTRUCTION>
You are an expert in **assessing the difficulty and medical knowledge requirements of medical image segmentation tasks**. Given a query image and its segmentation task, use the current diagnosis and relevant interaction history when available to assess whether additional task-specific anatomical knowledge is needed, and decide whether additional knowledge should be retrieved from an annotated image-mask example before generation.



<DYNAMIC CONTEXT>
Task Description | Compact Structured State | Input Image

<OUPUT>
{"need_support": *<true/false>*,
"reason": "*<brief justification>*"}

Figure 8: Knowledge Sufficiency Assessment prompt. The controller assesses whether additional task-specific anatomical knowledge is required from the task description, query image, and compact structured state, and returns a fixed JSON decision indicating whether support retrieval is needed.

in Fig. 8, the prompt receives the task description, query image, and compact structured state, and outputs a structured decision indicating whether an additional annotated support example should be retrieved. The retriever itself is feature-based and does not use a language-model prompt.

Generation Instruction Synthesis. When GENERATE is selected, the controller converts the task specification and relevant context into an executable instruction for the frozen UMM. As shown in Fig. 9, the prompt may additionally include a retrieved image–mask support pair. Retrieved examples are used only to provide semantic, anatomical, and appearance cues, rather than as geometric templates to be copied. The controller outputs the synthesized generation instruction in a fixed JSON format.

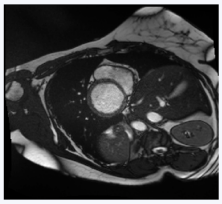
Editing Instruction Synthesis. When the retained prediction provides a reliable basis for local correction, the controller synthesizes a targeted editing instruction. As shown in Fig. 10, the controller reasons over the query image, retained best mask, residual-error diagnosis, and relevant editing history to determine what structure should be corrected, how it should be modified, and which reliable regions should be preserved. Previously ineffective corrections are also considered to avoid repeating failed refinement strategies.

Generation Execution. The synthesized generation instruction is then executed by the frozen UMM. As illustrated in Fig. 11, the image model receives the query image, the synthesized instruction, an optional selected support pair, and the dataset-specific output contract. It is instructed to produce only a spatially aligned RGB segmentation label map following the predefined label-color convention.

Editing Execution. Editing uses the same frozen UMM but a different visual context. As shown in Fig. 12, the image model receives only the query image, retained best mask, targeted editing

Generation Instruction-Synthesis Prompt

<STATIC INSTRUCTION>
You are an expert in **converting medical segmentation requirements and available context into an executable instruction for an image generation model**. Given the task, query image, relevant structured state, and retrieved support when available, synthesize a concise instruction that specifies the target structures, task-relevant anatomical cues, and required segmentation output. Use retrieved examples only as semantic and anatomical references, without copying their exact geometry, position, size, or contour.



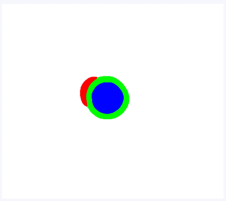
<DYNAMIC CONTEXT>
Task Description | Relevant State | Retrieved Support (Optional) | Input Image

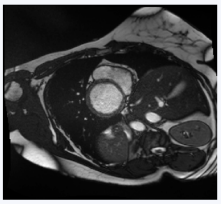
<OUTPUT FORMAT>
{"instruction": "<executable generation instruction>"}

Figure 9: Generation Instruction-Synthesis prompt. The controller combines the task description, stage-relevant structured state, query image, and optional retrieved support to produce an executable generation instruction for the frozen UMM.

Editing Instruction-Synthesis Prompt

<STATIC INSTRUCTION>
You are an expert in **translating residual segmentation errors into precise correction instructions for an image editing model.** Given the current best segmentation, its error diagnosis, and relevant interaction history, synthesize a targeted instruction that specifies what structure and region should be corrected, how it should be modified, and which reliable regions should be preserved. Avoid repeating previously ineffective corrections and choose an alternative refinement strategy when necessary.





<DYNAMIC CONTEXT>
Task Description | Compact Structured State | Current Best Mask | Input Image

<OUTPUT FORMAT>
{"target_label": "<label>",
"target_error": "<error type>",
"edit_instruction": "<targeted correction instruction>"}

Figure 10: Editing Instruction-Synthesis prompt. The controller transforms the retained prediction, residual-error diagnosis, and relevant interaction trajectory into a targeted correction instruction specifying the structure, error, and desired local revision.

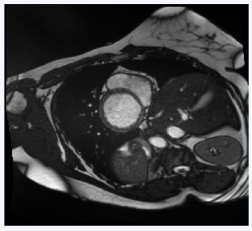
instruction, and output contract. Structured interaction trajectory and retrieved supports are not directly exposed to the editor; their relevant information has already been incorporated into the synthesized correction instruction.

Comparative Quality Critique. As illustrated in Fig. 13, the critic performs pairwise visual comparison rather than independent absolute scoring. Given the query image and two competing predictions, it compares segmentation quality label by label according to target identity, location, spatial extent, shape, topology, and boundary alignment. Optional retrieved supports may be provided only to clarify target semantics and coarse anatomy. The critic returns a structured preference together with label-wise error diagnosis, but does not determine the next inference action. Sequential pairwise comparisons are used to obtain the current tournament winner.

Adaptive Action Selection. Finally, the controller determines the next inference action from the current visual evaluation and interaction state. As shown in Fig. 14, ACCEPT is selected when all target labels are sufficiently reliable, EDIT when the retained prediction supports localized correction, and GENERATE when the current result remains globally unreliable or further local refinement

Generation Execution Prompt

<STATIC INSTRUCTION>
You are an image generation model performing medical image segmentation. Generate an RGB segmentation label map for the input image according to the provided instruction and output contract. The prediction should be anatomically consistent with and spatially aligned to the input. When a support example is provided, use it only as an anatomical and visual reference rather than copying its geometry. Output only the segmentation mask, without text, legends, overlays, borders, or other content.
<DYNAMIC CONTEXT>

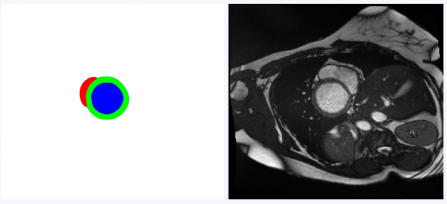


Generation Instruction | Selected Support Pair (Optional) | Input Image

Figure 11: Generation Execution prompt. The frozen UMM receives the query image, synthesized generation instruction, optional selected support pair, and output contract, and directly generates an RGB segmentation mask.

Editing Execution Prompt

<STATIC INSTRUCTION>
You are an image editing model performing local medical segmentation refinement. Given a query image, the current segmentation mask, and a correction instruction, revise only the specified structure or region while preserving reliable regions. The refined mask should remain anatomically consistent and spatially aligned with the query image, and follow the specified label definitions and RGB convention. Output only the final RGB segmentation mask, without text, legends, overlays, borders, or other content.
<DYNAMIC CONTEXT>



Editing Instruction | Current Best Mask | Input Image

Figure 12: Editing Execution prompt. The frozen UMM locally revises the retained best mask using the query image and synthesized correction instruction while preserving reliable regions and following the output contract.

is unlikely to succeed. When EDIT is selected, the controller additionally identifies the residual error to be targeted in the next iteration.

B.5 ADDITIONAL ANALYSES FOR OBSERVATIONS

To further support the observations in the main text, we provide two additional analyses on inference-time exploration and refinement, as summarized in Fig. 15. Fig. 15(a) analyzes diversity-aware retrieval and shows that varying retrieved supports further broadens the accessible hypothesis space. Fig. 15(b) examines the complementary roles of editing and resampling in prediction refinement.

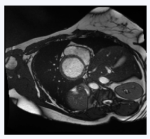
Diverse retrieval further broadens the accessible hypothesis space. Repeated sampling explores hypotheses within a fixed conditional context, while the context itself provides an additional source of diversity. We therefore diversify inference by conditioning generation on different retrieved supports S_i, such that $M_i \sim p_\theta(M \mid I, T, S_i)$. As shown in Fig. 15(a), under the same generation budget, distributing generations across diverse supports yields stronger Best-of-N performance than repeatedly sampling from a fixed support.

We further investigate how to retrieve complementary supports by comparing similarity-based retrieval with Maximal Marginal Relevance (MMR). MMR jointly considers query relevance and support redundancy, selecting the next support according to

$$S_i^* = \arg \max_{S \in \mathcal{D}_s \setminus \mathcal{S}} \left[\lambda \, \mathrm{sim}(I, S) - (1 - \lambda) \max_{S' \in \mathcal{S}} \mathrm{sim}(S, S') \right], \tag{8}$$

Comparative Quality Critique Prompt

<STATIC INSTRUCTION>
You are an expert in comparative medical image segmentation evaluation. Given a query image and two competing segmentations, compare them label by label based on target identity, location, spatial extent, anatomical shape, topology, and boundary alignment. Identify visible errors in both candidates, select the better prediction for each label, and describe the main residual error in the selected result. Retrieved supports, when provided, should be used only to clarify target semantics and coarse anatomy. Your role is limited to comparison and error diagnosis, not action selection.
<DYNAMIC CONTEXT>





Candidate A



Candidate B

Task Description | Candidate A | Candidate B | Input Image | Selected Support (Optional)

Figure 13: Comparative Quality Critique prompt. A frozen multimodal critic compares a baseline or current champion against a challenger, performs label-wise quality assessment and residual-error diagnosis, and returns the comparison in a fixed structured format.

Action Selection Prompt

<STATIC INSTRUCTION>
You are an expert in **planning iterative medical image segmentation**. Given the current visual evaluation, residual errors, and interaction state, select the next action from **GENERATE, EDIT, and ACCEPT**. Use ACCEPT when all target labels are sufficiently reliable, EDIT when the retained prediction provides a reliable basis for correcting a localized residual error, and GENERATE when the prediction remains globally unreliable or local refinement is unlikely to succeed. When selecting EDIT, identify one specific residual error to refine.
<DYNAMIC CONTEXT>

Task Description | Compact Structured State | Critic Evaluation

<OUTPUT FORMAT>
{"action": "<GENERATE | EDIT | ACCEPT>",
"target_label": "<label or null>",
"target_error": "<residual error or null>",
"reason": "<brief decision rationale>"}

Figure 14: Action Selection prompt. Based on the compact structured state and comparative critique, the controller selects GENERATE, EDIT, or ACCEPT, and specifies the target residual error when editing is chosen.

where $\mathcal{S}$ denotes the already selected supports and λ controls the trade-off between relevance and diversity. Fig. 15(a) shows that similarity-only retrieval tends to achieve slightly higher query relevance but also higher mutual redundancy, whereas MMR produces supports that are less redundant while remaining highly relevant. Specifically, MMR reduces the average support redundancy by 0.027 while sacrificing only 0.009 in query–support similarity, and improves the average Top-4 mDice from 60.04% to 62.26%.

To further quantify hypothesis diversity, we measure the standard deviation of Dice scores across sampled predictions. As shown in the inset of Fig. 15(a), hypothesis dispersion is positively correlated with Best-of-4 improvement ($\rho = 0.67$), suggesting that broader hypothesis exploration increases the chance of uncovering stronger predictions. Overall, diversity-aware retrieval provides relevant yet less redundant contexts, enabling more effective exploration of the hypothesis space under a fixed inference budget.

Editing and resampling offer complementary mechanisms for prediction refinement. Editing exploits an existing promising hypothesis through targeted correction, but is less effective for fundamental prediction errors, whereas resampling explores alternative hypotheses without explicitly preserving useful structures in the current prediction. We therefore investigate whether combining targeted editing with rollback and resampling can improve refinement robustness.

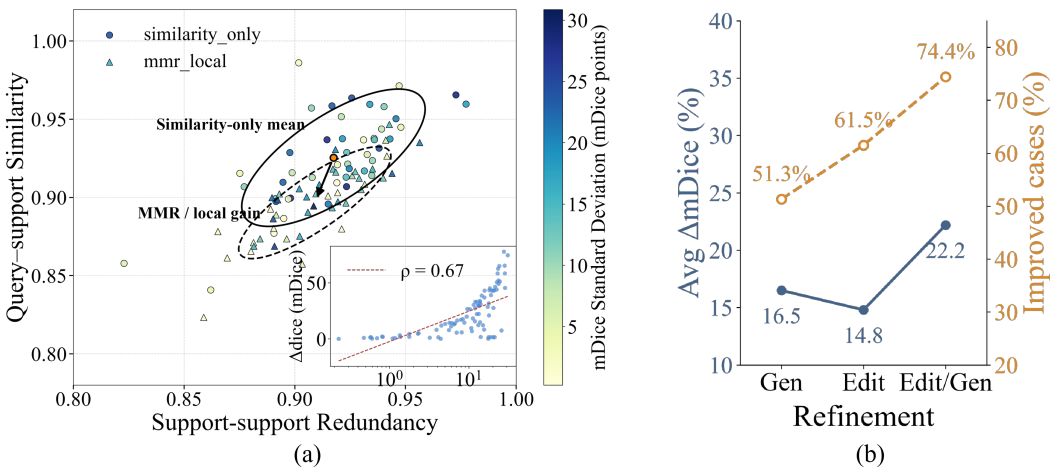


Figure 15: Additional analyses for the observations in the main text. (a) Diversity-aware retrieval broadens the accessible hypothesis space: query–support similarity and support redundancy are compared between similarity-only and MMR-based retrieval; marker color indicates hypothesis dispersion measured by the standard deviation of Dice scores across sampled predictions, and the inset shows its correlation with Best-of-4 improvement. (b) Editing and resampling provide complementary refinement mechanisms: the blue curve reports the mean Dice improvement and the orange curve reports the proportion of improved cases under regeneration, targeted editing, and their combination.

Table 7: Average inference cost of SegBanana across datasets. We report the average number of inference iterations, Nano Banana calls, and retrieval calls per sample.

Dataset	Iter.	Nano Banana Calls	Ret. Calls
ACDC	2.47	4.52	3.88
BraTS	1.76	3.40	2.26
BUS-UCLM	1.60	3.18	1.78
Drishti-GS	2.40	4.76	3.28
ISIC	1.70	3.24	1.52
Kvasir	1.56	3.12	0.92
RAVIR	2.25	4.50	3.33
TNBC	2.36	4.72	2.72
Overall	**1.94**	**3.75**	**2.32**

As shown in Fig. 15(b), generation alone yields an average improvement of 16.5 mDice points and improves 51.3% of cases. Targeted editing yields an average improvement of 14.8 points and improves 61.5% of cases. Combining editing with regeneration further increases the average improvement to 22.2 points and the proportion of improved cases to 74.4%. These results demonstrate the complementary roles of editing and resampling: editing enables targeted refinement of an existing hypothesis, while regeneration expands the hypothesis space when local correction is insufficient. Their combination provides the strongest overall refinement.

B.6 Agent Behavior and Inference Efficiency

Although SegBanana allows up to three inference iterations, its actual computation is dynamically determined by the evolving prediction state. We therefore analyze inference efficiency from two complementary perspectives: the average computational cost and the adaptive termination, action selection, and retrieval behavior of the controller.

Overall Inference Cost. Tab. 7 reports the average inference cost across datasets. Overall, SegBanana requires only 1.94 inference iterations, 3.75 valid Nano Banana calls, and 2.32 retrieval calls

Table 8: Adaptive stopping, action selection, and retrieval behavior of SegBanana. Stop@k denotes the percentage of samples terminating after iteration k. Action columns report the distribution of controller decisions after capping GENERATE to at most two actions per iteration, and Retrieved denotes the percentage of samples invoking support retrieval at least once.

Dataset	Stop@1	Stop@2	Stop@3	Generate	Edit	Accept	Retrieved
ACDC	15.0%	23.3%	61.7%	76.6%	14.4%	9.0%	100.0%
BraTS	58.0%	8.0%	34.0%	63.5%	15.3%	21.2%	96.0%
BUS-UCLM	66.0%	8.0%	26.0%	55.8%	17.6%	26.7%	90.0%
Drishti-GS	28.0%	4.0%	68.0%	65.3%	24.8%	9.9%	100.0%
ISIC	58.0%	14.0%	28.0%	56.5%	20.6%	22.9%	62.0%
Kvasir	64.0%	16.0%	20.0%	52.5%	20.3%	27.2%	44.0%
RAVIR	33.3%	8.3%	58.3%	67.3%	16.4%	16.4%	100.0%
TNBC	20.0%	24.0%	56.0%	57.1%	32.8%	10.1%	100.0%
Overall	**46.0%**	**14.0%**	**40.1%**	–	–	–	**83.2%**

per sample. Here, valid Nano Banana calls count successful generation outputs, with at most two counted per iteration, matching the inference-time generation budget.

The inference cost varies substantially across datasets. Kvasir and BUS-UCLM require only 1.56 and 1.60 iterations on average, whereas ACDC and Drishti-GS require 2.47 and 2.40, respectively. A similar pattern appears in visual-model usage: Kvasir requires only 3.12 valid Nano Banana calls per sample, while Drishti-GS and TNBC require 4.76 and 4.72. This variation shows that SegBanana does not follow a fixed execution trajectory, but allocates additional inference when the current prediction remains difficult to resolve.

Adaptive Termination. Tab. 8 further summarizes the termination behavior. Overall, 46.0% of cases terminate after the first iteration and 14.0% after the second, while 40.1% reach the third iteration. Thus, 59.9% of cases terminate before exhausting the maximum inference budget.

The stopping distribution is strongly task dependent. Most BUS-UCLM, Kvasir, BraTS, and ISIC cases terminate after the first iteration, with Stop@1 rates of 66.0%, 64.0%, 58.0%, and 58.0%, respectively. In contrast, Drishti-GS, ACDC, RAVIR, and TNBC more frequently require the full three-iteration budget, reaching Stop@3 rates of 68.0%, 61.7%, 58.3%, and 56.0%. These results indicate that the controller adapts not only the operation performed within each iteration but also the overall inference depth.

Adaptive Action and Retrieval Usage. Generation remains the dominant action across all datasets, accounting for 52.5%–76.6% of controller decisions. Editing is used more selectively, ranging from 14.4% on ACDC to 32.8% on TNBC, consistent with its role in locally refining predictions that already provide a reliable basis for correction. The proportion of ACCEPT decisions also varies markedly across tasks, reflecting different stopping behaviors.

Knowledge retrieval is similarly adaptive rather than uniformly applied. Overall, 83.2% of cases invoke retrieval at least once. Retrieval is used for all cases on ACDC, Drishti-GS, RAVIR, and TNBC, but only 44.0% of Kvasir and 62.0% of ISIC cases require external support. This variation shows that the controller acquires additional anatomical context according to the task and evolving interaction state rather than enforcing retrieval for every sample.

B.7 ADDITIONAL CASE STUDIES

Figs. 16 and 17 provide additional qualitative comparisons across eight datasets. SegBanana produces predictions that remain visually consistent with the target anatomy across substantially different modalities and structural characteristics. For relatively compact targets such as breast lesions, polyps, skin lesions, and brain tumors, its predictions generally preserve the overall target extent and shape, whereas several baselines exhibit substantial under-segmentation, over-segmentation, or foreground–background confusion. For more structurally demanding tasks, the differences become more apparent: SegBanana better preserves the nested disc–cup relationship in Drishti-GS and the

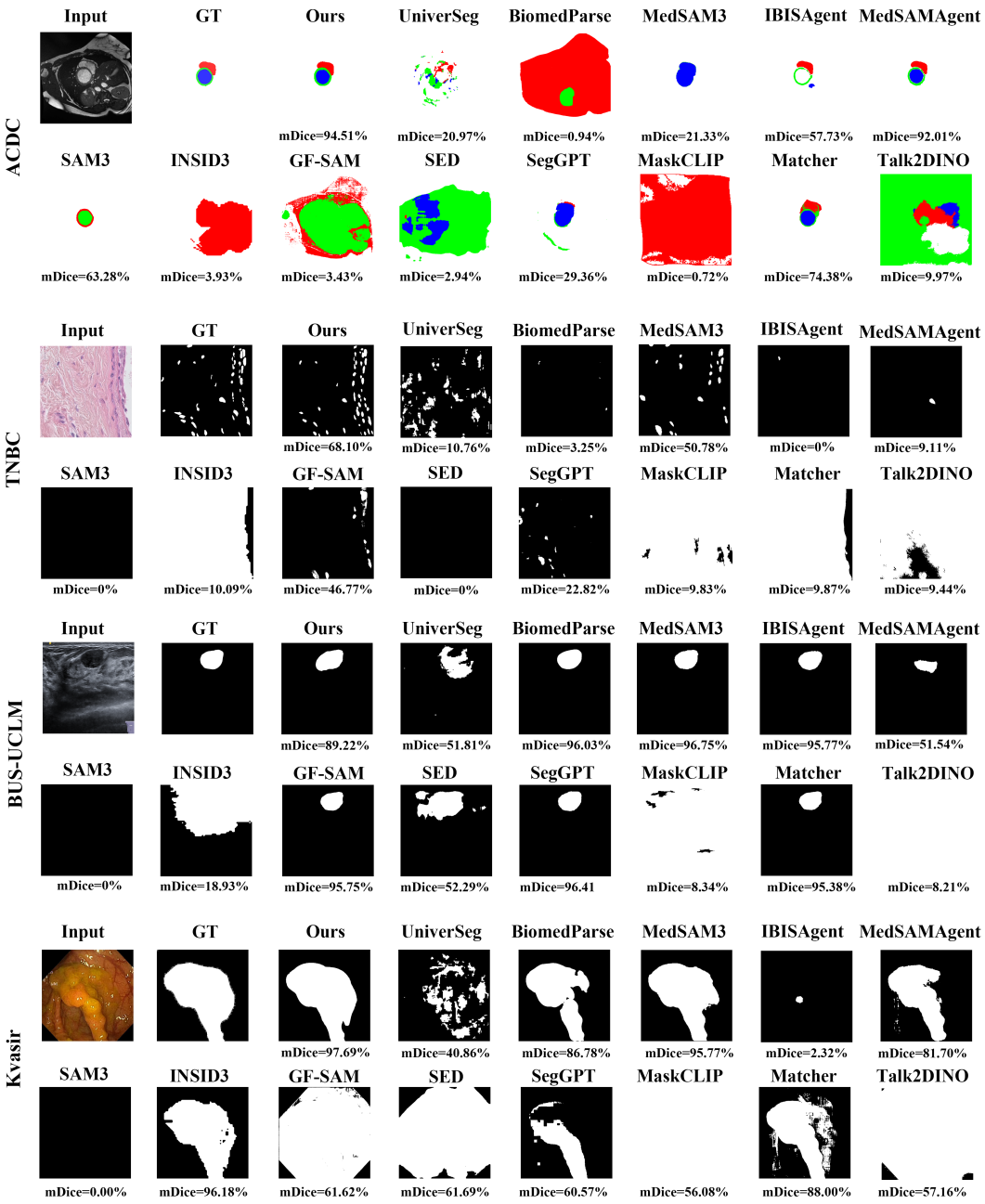


Figure 16: Additional qualitative comparisons on ACDC, TNBC, BUS-UCLM, and Kvasir. We compare SegBanana with representative medical-specific and generalist segmentation methods across cardiac MRI, histopathology, ultrasound, and endoscopy.

thin, branching vascular structures in RAVIR, while several competing methods collapse these structures into coarse regions or miss large portions of the target.

The histopathology and cardiac examples further illustrate the challenge of transferring across domains. On TNBC, several methods produce sparse, fragmented, or nearly empty masks, whereas SegBanana captures a denser distribution of nuclei consistent with the reference. On ACDC, SegBanana better preserves the relative spatial arrangement of multiple cardiac structures, while some

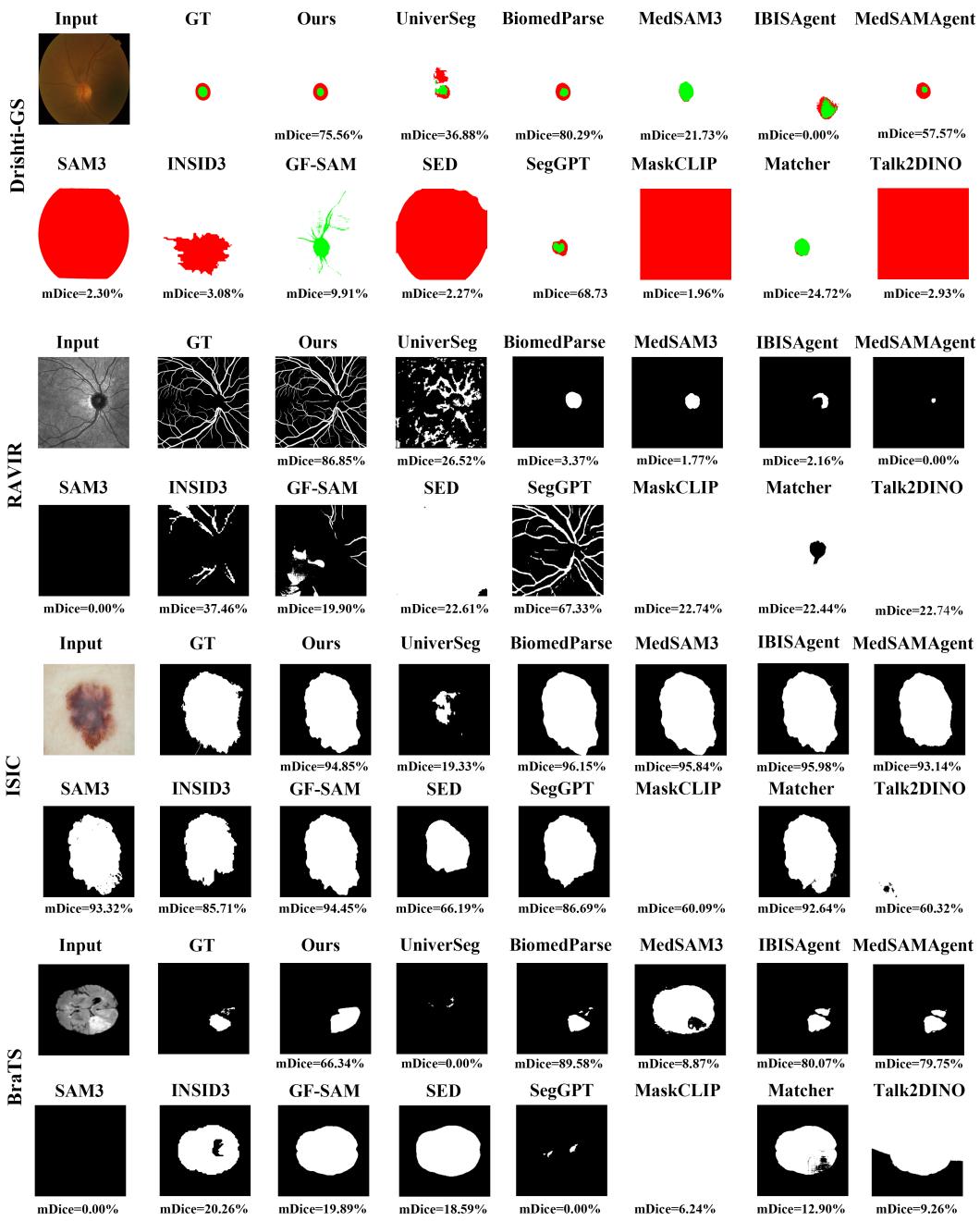


Figure 17: Additional qualitative comparisons on Drishti-GS, RAVIR, ISIC, and BraTS. The examples cover nested anatomical structures, fine vascular structures, skin lesions, and brain tumors, illustrating segmentation behavior across diverse targets and imaging modalities.

baselines confuse labels or substantially distort their extent. Overall, these examples qualitatively support the broad transferability of SegBanana across targets ranging from compact lesions to fine and multi-structure anatomy.