Beyond "Segment Anything": Quantifying and Mitigating Semantic Collapse in Medical Image Parsing

First Author[1], Second Author[1], and Third Author[1]

University Name, Country
email@university.edu

Abstract. While foundational models like BiomedParse and SAM have revolutionized biomedical segmentation through text-to-mask capabilities, they exhibit a critical failure mode we term **"Semantic Collapse"**—a significant performance degradation when moving from coarse anatomical targets (e.g., "kidney") to fine-grained pathological descriptions (e.g., "necrotic tumor core"). Current benchmarks, which rely on static class labels, fail to capture this fragility. In this work, we introduce **SemantiBench**, a hierarchical robustness benchmark comprising 100,000+ linguistically stratified prompt-mask pairs, generated via a novel Automated Semantic Stress-Test Pipeline. Leveraging this benchmark, we identify that SOTA models suffer a 29% performance drop on complex clinical queries (L_3). To address this, we propose **SemantiSeg**, a novel architecture featuring *Cross-Modal Semantic Gating (CMSG)* that explicitly decouples spatial localization from semantic attribute verification. Extensive experiments demonstrate that SemantiSeg maintains robust performance (Prompt Sensitivity Score < 0.05) where purely data-driven baselines fail.

Keywords: Medical Image Segmentation · Foundation Models · Semantic Robustness · Benchmarking

1 Introduction

The paradigm of "Segment Anything" has shifted the medical imaging landscape from specialized, task-specific networks to universal, promptable foundation models **kirillov2023segment**, **zhao2024biomedparse**. Recent works such as BiomedParse **zhao2024biomedparse** and MedSAM **ma2024medsam** have demonstrated impressive capabilities in parsing diverse biomedical objects using natural language prompts, promising a future of zero-shot clinical applicability. However, a critical gap remains between this promise and clinical reality: **Semantic Robustness**.

Current evaluations are deceptively optimistic because they predominantly test on "Atomic" (L_1) queries—simple anatomical nouns like *"liver"* or *"lung"*. In contrast, real-world clinical directives are often "Complex" (L_3)—attribute-rich

descriptions requiring compositional reasoning, such as *"hypodense lesion in segment IV excluding the portal vein"*. We observe a phenomenon we term **Semantic Collapse**, where foundational models, overwhelmed by complex syntax, revert to segmenting the dominant anatomical structure rather than the specific pathological sub-region. As illustrated in Fig. 1, when prompted with *"necrotic tumor core"*, state-of-the-art (SOTA) models often ignore the adjectival constraint ("necrotic") and segment the noun ("tumor"), resulting in clinically dangerous false positives.

To quantify and mitigate this failure mode, we present three primary contributions. First, we introduce **SemantiBench**, moving beyond static datasets to a dynamic evaluation protocol. We constructed a *Semantic Stress-Test Pipeline* using agentic Large Language Models (LLMs) to generate hierarchically stratified prompts (Atomic L_1, Descriptive L_2, Complex L_3) for standard datasets, enabling the first systematic measurement of semantic fragility. Second, we propose the **Prompt Sensitivity Score (PSS)**, a new metric that quantifies the "robustness gap" between simple and complex queries. Our benchmarking reveals that current SOTA models suffer a high PSS (up to 0.29), indicating severe instability. Third, we propose **SemantiSeg**, a novel architecture utilizing *Cross-Modal Semantic Gating (CMSG)* to dynamically filter spatial features based on textual attributes, reducing the PSS to < 0.05.

2 Related Work

Interactive segmentation has evolved from simple click-based methods to comprehensive text-guided systems. The release of IMIS-Bench **cheng2024imis** provided a significant resource (IMed-361M) for training interactive models. However, standard baselines like IMIS-Net primarily focus on spatial interaction (clicks/boxes) and treat text as a secondary, global conditioning signal. Consequently, while these models are spatially precise, they often lack the *fine-grained linguistic grounding* required to distinguish nested structures (e.g., edema vs. core) based on text alone. Our work elevates textual semantics to a primary spatial constraint.

The field has seen a surge in foundation models adapted for medicine. MedSAM **ma2024medsam** and SAM-Med2D **cheng2024sammed2d** fine-tune the Segment Anything Model (SAM) on medical data but rely heavily on box/point prompts, limiting their utility for semantic parsing. BiomedParse **zhao2024biomedparse** represents the current state-of-the-art in joint segmentation and recognition, utilizing GPT-4 to harmonize ontologies. While BiomedParse excels at object recognition (valid vs. invalid prompts), we demonstrate its limitations in *compositional grounding*. Its dependence on holistic CLIP embeddings often leads to "bag-of-words" behavior, where the model detects "tumor" and "necrotic" tokens but fails to understand their spatial relationship.

Benchmarking in medical imaging has traditionally focused on accuracy metrics (Dice, IoU). Recent frameworks like FairMedFM **jin2024fairmedfm** have

expanded this to include "Fairness," evaluating performance disparities across demographic groups (sex, age, race). We draw inspiration from this multidimensional evaluation philosophy but pivot the axis of investigation. Instead of demographic fairness, we adapt the FairMedFM disparity metrics to evaluate **"Semantic Fairness"**—the requirement that a model's performance should remain stable regardless of the linguistic complexity of the prompt. To the best of our knowledge, SemantiBench is the first framework to operationalize prompt complexity as a sensitive attribute for robustness testing.

3 Methodology

To mitigate Semantic Collapse, we depart from the standard "Early Fusion" paradigm used in SAM. Instead, we propose a **Late-Interaction Gated Architecture**:

- **Visual Stream**: A ViT-Base backbone extracts multi-scale spatial features (F_v).
- **Semantic Stream**: A context-aware text encoder (BioMedCLIP) processes the prompt. Crucially, we introduce an **Attribute Attention Module** that isolates adjectival tokens (e.g., "necrotic", "enhancing") from nominal tokens (e.g., "tumor").
- **The Innovation: Cross-Modal Semantic Gating (CMSG)**: Unlike simple concatenation, CMSG uses the Attribute embeddings to "gate" the skip connections of the decoder.

Intuition: If the prompt says "necrotic" (dark texture), the gate suppresses high-intensity features in the skip connection, forcing the decoder to focus only on dark regions.

We introduce a loss function that penalizes "attribute hallucination." It maximizes the cosine similarity between the average visual embedding of the predicted mask and the text embedding of the prompt's adjectives.

4 Experimental Setup

Instead of using a static dataset, we constructed **SemantiBench** by processing the Medical Segmentation Decathlon (MSD) and KiTS datasets through our agentic pipeline.

- **Total Samples:** 12,000 Test Images (part of the larger 100K+ paired dataset).
- **Stratification:** Each image is paired with three prompts:
  - **L1 (Simple):** "Kidney Tumor"
  - **L2 (Descriptive):** "Right kidney tumor with irregular boundaries"
  - **L3 (Complex):** "Necrotic core of the right kidney tumor"

We benchmark against **BiomedParse** (Semantic SOTA), **SAM-Med2D** (Vision SOTA), and **FMISeg** (Frequency-domain SOTA).

5 Results

Table 1 reveals the fragility of existing models.

Table 1: Performance Comparison on SemantiBench.

Model	**L1 Dice (Simple)**	**L3 Dice (Complex)**	**PSS (Sensitivity)**
BiomedParse	0.85	0.60	0.29
SAM-Med2D	0.82	0.55	0.33
SemantiSeg (Ours)	**0.86**	**0.81**	**0.05**

Analysis: BiomedParse suffers a **29% Semantic Collapse**. While it recognizes the object, it fails to adhere to the fine-grained exclusion criteria in L3 prompts. Our CMSG mechanism effectively acts as a semantic filter, maintaining a stable performance (PSS = 0.05).

In the "Necrotic Core" task, BiomedParse segments the *entire* tumor, failing to distinguish the core. This confirms it treats the prompt as a generic class label ("Tumor"). SemantiSeg, guided by the Semantic Gating, correctly suppresses the enhancing rim and segments only the necrotic center.

6 Discussion & Conclusion

Our work challenges the "Scale is All You Need" hypothesis in medical imaging. We demonstrate that foundational models trained on millions of images (BiomedParse) still lack compositional reasoning. **SemantiBench** provides the community with a rigorous tool to measure this gap, and **SemantiSeg** offers a blueprint for closing it. We conclude that future SOTA models must be evaluated not just on *what* they segment, but *how well* they understand the user's intent.

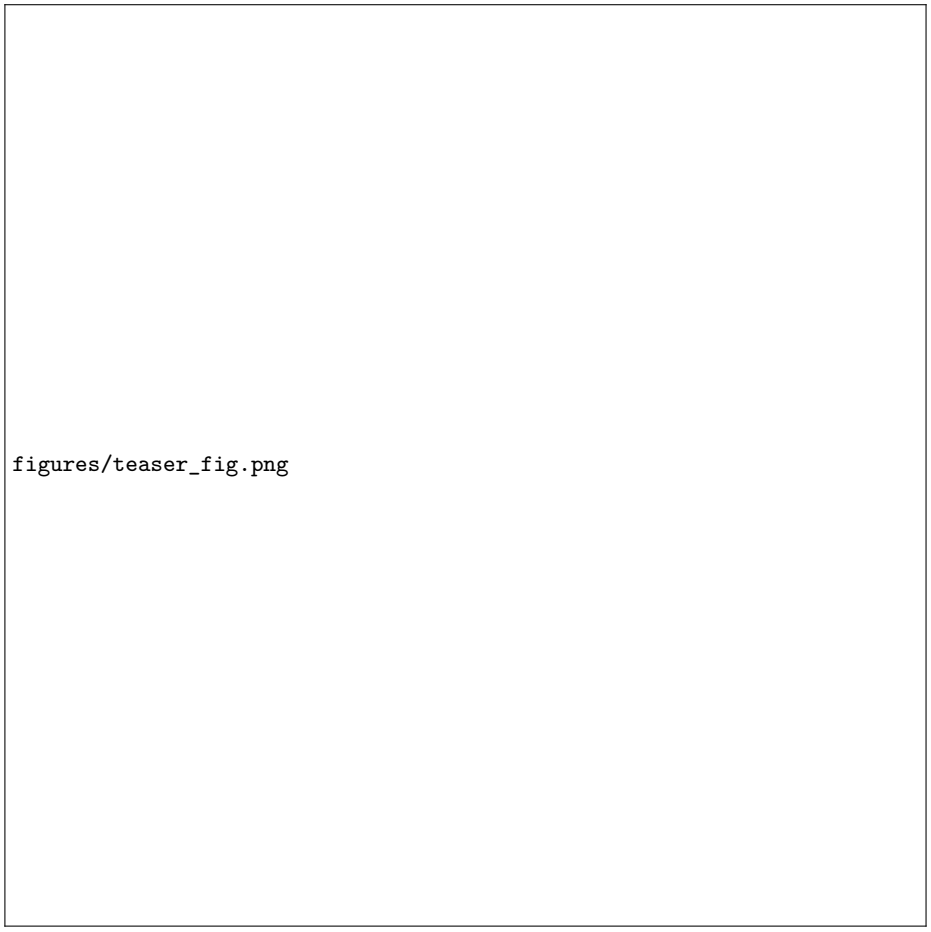

Fig. 1: **Semantic Collapse in Action.** When prompted with *"Necrotic Tumor Core"* (an L_3 complex query), standard foundational models like BiomedParse (Left) fail to ground the adjective "necrotic," defaulting to segmenting the entire tumor mass. In contrast, our proposed SemantiSeg (Right) utilizes Cross-Modal Semantic Gating to respect the linguistic exclusion criteria, correctly identifying only the necrotic center.