SemantiBench: Measuring Prompt Compliance in Medical Image Segmentation with Frequency-Gated Models

Ngo Thanh Trung[1⋆], Tran Long Vu[1⋆⋆], and Doan Hoang Long[1⋆⋆⋆]

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
{Trung.NT227990, Vu.TL228054, Long.DH228027}@sis.hust.edu.vn

Abstract. Foundation models like BiomedParse and SAM have improved biomedical segmentation with text-to-mask capabilities, but they often fail when prompts shift from simple anatomical targets (e.g., "kidney") to fine-grained pathological descriptions (e.g., "necrotic tumor core"). We refer to this degradation as *semantic collapse*. We identify that this stems from CLIP's inability to represent negation geometrically—prompts like "excluding tumor" still activate tumor-related features. In this paper, we introduce **SemantiBench**, a dataset of 100,000+ prompt-mask pairs, and propose **FreqMedCLIP**, a segmentation model that explicitly separates target and avoidance signals. Unlike standard methods, our model decouples prompts into Target and Avoidance streams and enforces constraints via a **logical gate**. We also introduce an **exclusion loss** that supervises the gate to act as a precise detector for forbidden regions. Our model achieves a Dice score of 0.77 on complex exclusionary queries (L_3), significantly outperforming the baseline (Dice 0.60). Current models fail this test; they ignore "excluding" and segment everything.

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

1 Introduction

Foundation models such as BiomedParse [16] and MedSAM [11] can now segment biomedical objects using natural language. However, strict adherence to clinical prompts remains a problem. While these models succeed at atomic queries (L_1) like "kidney", they often fail to comply with complex constraints (L_3) such as "kidney excluding the renal pelvis".

This failure often stems from "semantic collapse", where the model ignores key logical or descriptive modifiers and incorrectly defaults to the generic object definition. For example, when prompted with "necrotic tumor core", many state-of-the-art models ignore the adjective "necrotic" and segment the entire tumor. This behavior is dangerous in clinical settings where precise sub-region targeting is required.

This paper introduces three contributions to address this problem. First, we propose **SemantiBench**, a protocol to measure two distinct forms of robustness: (1) **Descriptive Invariance** (L_2), which checks if models output consistent masks for synonymous prompts (e.g., "kidney" vs. "bean-shaped organ"); and (2) **Logical Compliance** (L_3), which tests if models correctly modify the segmentation mask when prompts impose exclusionary constraints.

To quantify these failures, we introduce the **Prompt Sensitivity Score (PSS)**. We find that current models exhibit a PSS gap of up to 0.29, indicating poor compliance with complex instructions.

⋆ Equal contribution
⋆⋆ Equal contribution
⋆⋆⋆ Equal contribution

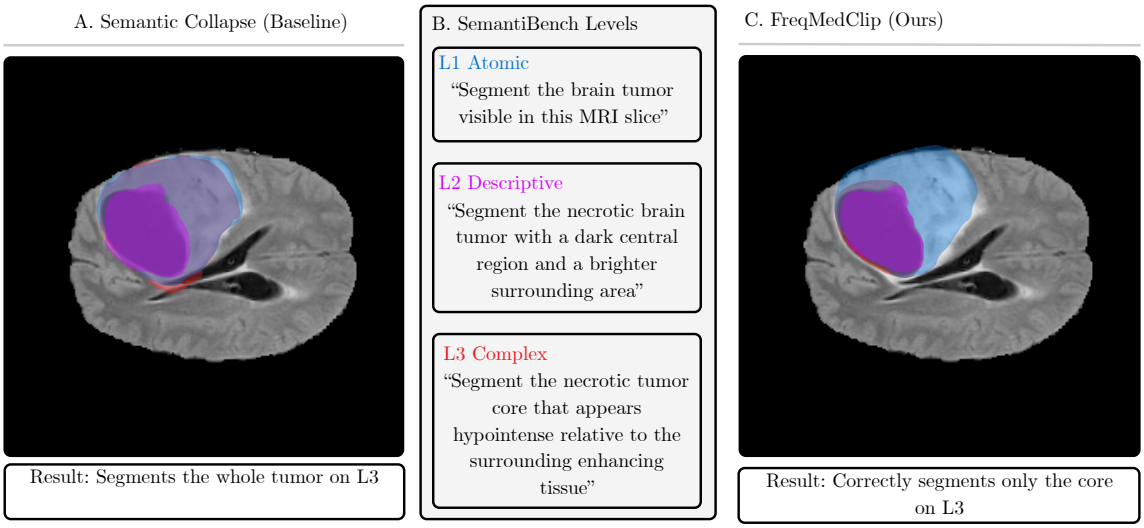

Fig. 1: **Semantic failure modes.** (A) Existing foundation models (BiomedParse) fail to process the adjective "necrotic," incorrectly segmenting the entire tumor mass [16]. (B) SemantiBench evaluates robustness across three linguistic levels ($L_1 - L_3$). (C) FreqMedClip uses cross-modal gating to isolate the necrotic core.

Finally, we present **FreqMedCLIP**, an architecture designed to solve this ambiguity. We argue that CLIP embeddings suffer from what we call "affirmative bias"—the encoder treats all words in a prompt as inclusion signals, even negation operators like "excluding." FreqMedCLIP addresses this by decoupling the text into Target and Avoidance streams. We introduce a **logical gating** mechanism that forces the exclusion through multiplicative gating, suppressing the forbidden region at the feature level. This approach, combined with explicit exclusion loss, reduces the PSS to 0.12, significantly improving prompt compliance.

2 Related Work

Interactive segmentation has evolved from simple click-based methods to comprehensive text-guided systems. The release of IMIS-Bench [1] 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 [11] and MedClipSamV2 [6] fine-tune the Segment Anything Model (SAM) on medical data but rely heavily on box/point prompts, limiting their utility for semantic parsing. BiomedParse [16] 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 [5] 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

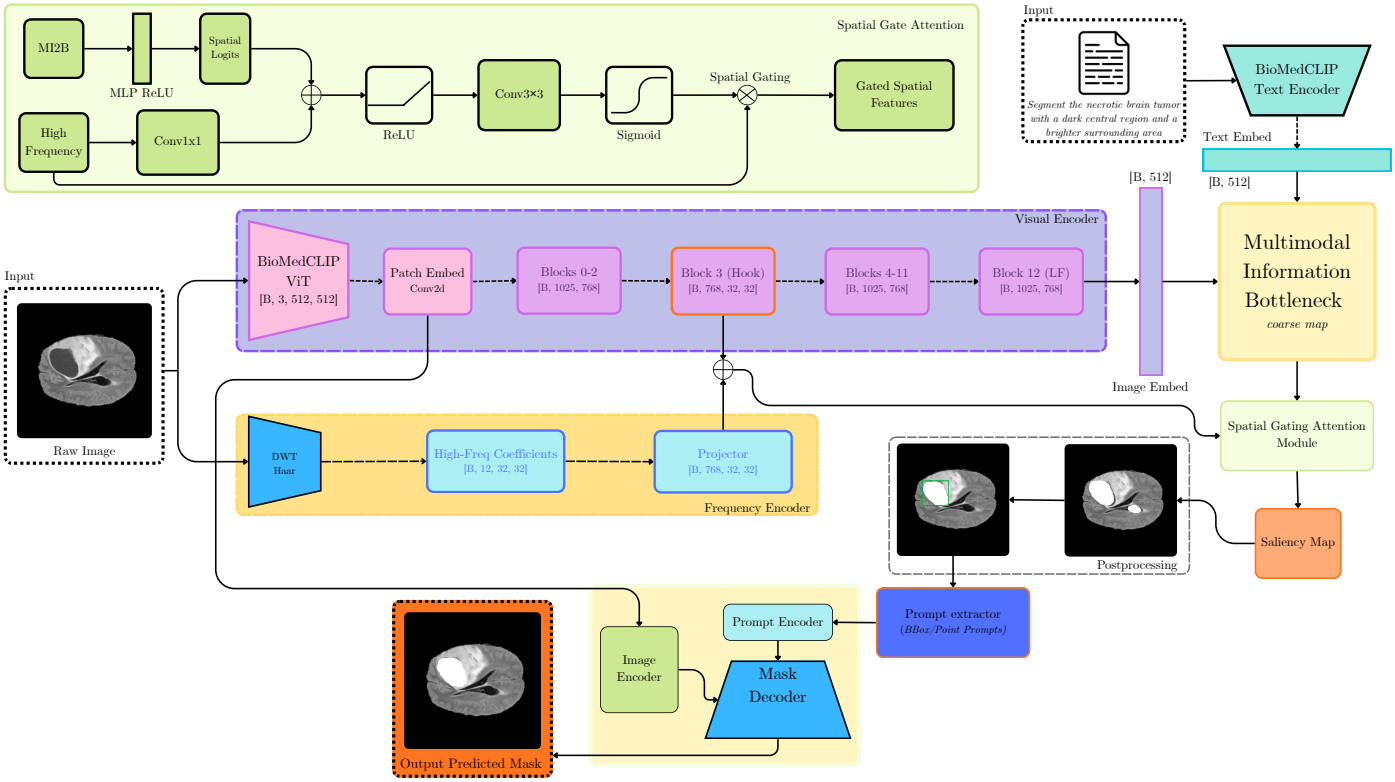

Fig. 2: **FreqMedClip Architecture.** The model uses two distinct streams: (1) A visual backbone (BiomedCLIP) that extracts global context, and (2) A High-Frequency Encoder that captures the fine texture and boundary details associated with specific attributes (e.g., necrotic, spiculated). The Cross-Modal Semantic Gating (CMSG) mechanism actively filters this high-frequency stream using the text prompt, isolating only the structures that match the logical query.

We designed FreqMedCLIP to bridge the Frequency-Semantic Gap in medical segmentation. We assume that critical semantic attributes, like whether a tumor is necrotic or nodular, are encoded in high-frequency signals

that standard Vision Transformers (ViTs) tend to smooth out. Consequently, our model treats the high-frequency stream not just as a simple edge detector, but as a rich source of raw material for semantic reasoning.

FreqMedClip: Frequency-Aware Logic.

High-Frequency Attribute Encoder. Standard ViTs suffer from a key limitation: they act as low-pass filters, effectively blinding them to texture-defined semantics [14]. While recent methods have tried to add frequency modeling [8,4], they often treat it as a secondary auxiliary task. We take a different approach by explicitly encoding the high-frequency spectrum as a primary input.

We use a **Discrete Wavelet Transform (DWT)** to decompose the input image $\mathbf{X}$ into four spectral sub-bands:

$$\mathbf{X}_{freq} = \text{DWT}(\mathbf{X}) = \{\mathbf{X}_{LL}, \mathbf{X}_{LH}, \mathbf{X}_{HL}, \mathbf{X}_{HH}\} \tag{1}$$

The components $\{\mathbf{X}_{LH}, \mathbf{X}_{HL}, \mathbf{X}_{HH}\}$ capture vertical, horizontal, and diagonal details respectively. These are precisely where features like tissue texture (e.g., necrotic vs. vital) and margin type (e.g., spiculated vs. smooth) reside. We feed these 12 spectral channels into a **ConvNeXt-Tiny** encoder. This encoder does not just extract edges. It produces the semantic raw material that our text gate will later act upon.

Dual-Stream Encoder. For global context, we use the **BiomedCLIP ViT-B/16**. To handle complex, compositional queries like "Kidney excluding tumor", we use a semantic decomposition strategy. We split the input prompt into two parts: P_{pos} (the target, e.g., "Kidney") and P_{neg} (the avoidance target, e.g., "Tumor"). We encode these into two distinct embedding vectors, $\mathbf{E}_{pos}$ and $\mathbf{E}_{neg}$. This distinction is critical because it ensures the avoidance signal remains independent from the target representation, preventing the model from confusing what it should finding with what it should ignore.

Logical Gating (CMSG). We propose a **Cross-Modal Semantic Gating (CMSG)** mechanism. Unlike standard cross-attention, which simply mixes signals based on similarity, our gate operates on strict exclusion logic. The high-frequency feature map $\mathbf{F}_{freq}$ passes through two parallel paths:

- **Inclusion Path ($\mathbf{G}_{inc}$):** This path highlights regions that match $\mathbf{E}_{pos}$.
- **Exclusion Path ($\mathbf{G}_{exc}$):** This path detects regions that match $\mathbf{E}_{neg}$.

The final refinement step enforces our logical constraint: we keep the regions matching P_{pos} but actively suppress those matching P_{neg}:

$$\mathbf{F}_{refined} = \mathbf{F}_{freq} \odot \mathbf{G}_{inc} \odot (1 - \mathbf{G}_{exc}) \tag{2}$$

This operator effectively turns off features that align with the forbidden concept, allowing the model to carve out the target object even when it overlaps with the excluded region.

Exclusion Loss. Implicit supervision (hoping the model figures out what not to segment) is insufficient for teaching strict negation. To fix this, we introduce an **exclusion loss** (L_{excl}). During training, we use multi-class ground truth to directly supervise the Exclusion Path ($\mathbf{G}_{exc}$). We force this path to detect the excluded class explicitly. This ensures that the suppression term $(1 - \mathbf{G}_{exc})$ receives strong, direct gradient signals.

$$L_{total} = L_{Dice}(\mathbf{Y}_{pred}, \mathbf{Y}_{target}) + \lambda L_{BCE}(\mathbf{G}_{exc}, \mathbf{Y}_{avoid}) \tag{3}$$

SemantiBench Existing benchmarks typically test object recognition using simple nouns (e.g., "Segment the Liver"). They fail to test if a model actually understands descriptive logic or negation. To rigorously evaluate this, we built **SemantiBench**, a dataset specifically designed to probe semantic compliance.

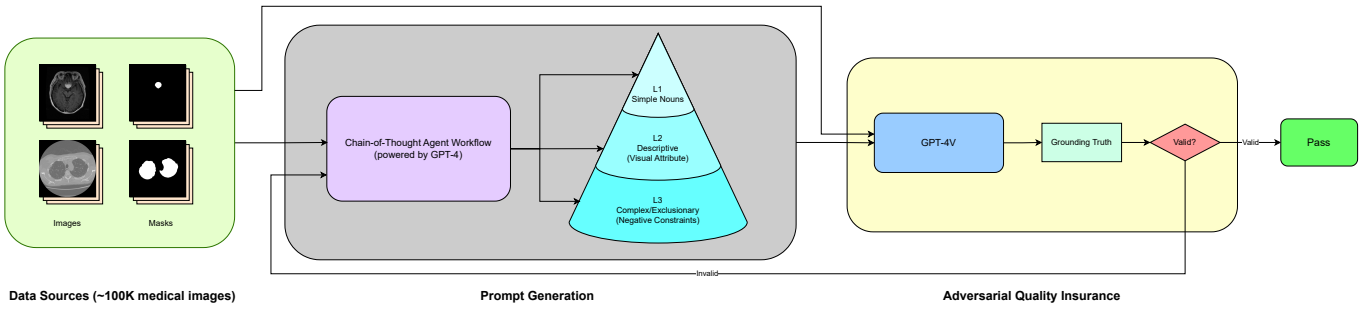

Fig. 3: **SemantiBench Construction Pipeline.** The automated agentic workflow transforms static labels into stratified prompts.

Prompt Generation Pipeline. We need to ensure that our Exclusionary (L_3) prompts map to ground-truth masks that are distinct from the atomic (L_1) masks. Handling label ambiguity is a known challenge in medical segmentation [15], especially for hierarchical structures like tumors. We use multi-label datasets (such as KiTS23 and MSD-Liver) where sub-regions are clearly annotated (e.g., Kidney=1, Tumor=2, Cyst=3). We map these to three distinct levels of complexity:

- L_1 **(Atomic):** "Kidney." The target is the Union of labels 1, 2, and 3. This tests basic recognition.
- L_2 **(Descriptive):** "The bean-shaped organ..." The target remains the Union(1, 2, 3). Since the target is identical to L_1, we use this to measure **Invariance**: does the model produce the same mask when the prompt changes from a name to a description?
- L_3 **(Compliance):** "Kidney excluding tumor." The target is now Label 1 only. Here, the target mask actually changes. We measure **Compliance**: can the model adjust its boundary based on the logical instruction?

This structured mapping prevents the ground truth paradox, where a model is penalized for correctly excluding a region that is present in the static mask of a generic dataset.

Verification Loop. Synthetic data generation entails a risk of hallucination (generating descriptions for attributes that do not exist in the specific image slice) [3]. To mitigate this, we implemented a strict verification loop. A Vision-Language Model (GPT-4V) acts as a quality assurance agent. It reviews each generated prompt against the image slice. The VLM verifies that the attribute described (e.g., cyst) is visibly present. If the attribute is missing or ambiguous, we discard the prompt from the benchmark. This ensures that SemantiBench measures the model's ability to segment, not its ability to guess.

4 Datasets and Experimental Setup

4.1. Prompt Generation Pipeline. To rigorously evaluate compliance, we must ensure that our "Exclusionary" (L_3) prompts map to ground-truth masks that are distinct from the atomic (L_1) masks. Handling label ambiguity is a known challenge in medical segmentation [15], particularly for hierarchical structures like tumors [13]. We utilize multi-label datasets (KiTS23, MSD-Liver) where sub-regions are annotated (e.g., Kidney=1, Tumor=2, Cyst=3).

- L_1 **(Atomic):** "Kidney." Target = Union(1, 2, 3).
- L_2 **(Descriptive):** "The bean-shaped organ..." Target = Union(1, 2, 3). Since the target is identical to L_1, we measure **Invariance** (Similarity between Model(L_1) and Model(L_2)).
- L_3 **(Compliance):** "Kidney excluding tumor." Target = Label 1 only. Here, the target mask changes. We measure **Compliance** (Dice between Model(L_3) and Label 1).

This mapping prevents the "ground truth paradox," where a model is penalized for correctly excluding a region present in the static mask.

4.2. Verification Loop. We maintain a verification loop to filter out hallucinations, a critical risk in synthetic medical data generation [3]. A VLM (GPT-4V) verifies that the attribute described in the prompt (e.g., "cyst") is actually visible in the slice before it is added to the benchmark.

4.3. Evaluation Metrics. Following recent guidelines on metric pitfalls [12] and hallucination benchmarking [9], we employ the Dice coefficient as our primary segmentation metric, along with robustness-focused derivatives.

Dice Score. The Dice coefficient measures the overlap between the predicted segmentation mask and the ground truth:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \tag{4}$$

where A is the predicted mask and B is the ground truth. We interpret the scores as follows: 1.0 indicates perfect alignment; 0.8–0.9 represents performance equivalent to inter-expert consensus in clinical settings [12]; and scores below 0.6 indicate substantial boundary errors or over-segmentation into irrelevant regions.

Building on this foundation, we define two robustness metrics:

1. **Prompt Sensitivity Score (PSS):** Measures the drop in performance when moving from L_1 to L_3. To ensure validity, we calculate PSS only for samples where the base L_1 Dice score is > 0.8.

$$PSS = 1 - \frac{\text{Dice}(L_3)}{\text{Dice}(L_1)} \quad (\text{Valid only if Dice}(L_1) > 0.8) \tag{5}$$

2. **L2 Invariance Similarity:** Dice between the prediction for the canonical name (L_1) and the prediction for a descriptive synonym (L_2), measuring whether the model can segment the same region when only descriptive text is provided. We compute it as:

$$\text{Dice}_{L2} = \frac{2\,|\text{pred}(L_1) \cap \text{pred}(L_2)|}{|\text{pred}(L_1)| + |\text{pred}(L_2)|} \tag{6}$$

4.4. Baselines. We compare against: (1) **BiomedParse** [16]: SOTA foundation model. (2) **SAM-Med2D** [2]: Adapter-based SAM. (3) **UNet-CLIP (Prompt Tuning)**: Tests if simple concatenation is sufficient [7]. (4) **LViT** [10]: A Language-Vision Transformer that uses text to modulate attention.

5 Results

Implementation Details. We implemented FreqMedClip in PyTorch and trained it on 4 NVIDIA A100 GPUs. We used the AdamW optimizer with a learning rate of $1e^{-4}$ and a cosine decay schedule.

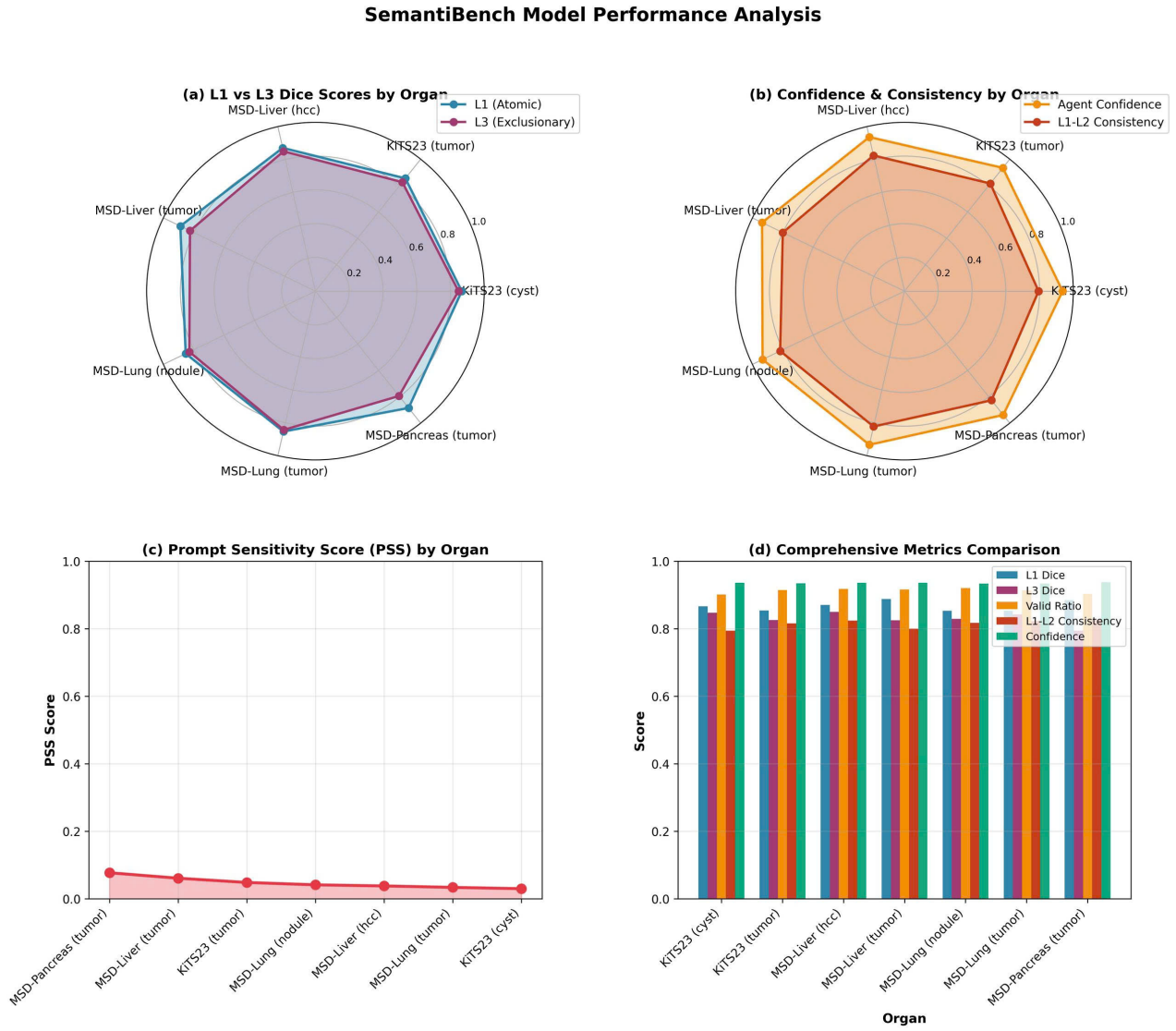

Fig. 4: **Quantitative analysis of the SemantiBench dataset characteristics and validation metrics across multi-organ pathologies. (a)** Radar plot comparing Atomic (L1) vs. Exclusionary (L3) Dice scores; the minimal overlap reduction indicates resistance to semantic collapse. **(b)** Correlation between Agent 2 (Vision Critic) confidence and L1-L2 descriptive consistency, validating the reliability of synthetic prompts. **(c)** Prompt Sensitivity Score (PSS) trend, demonstrating a low mean PSS (0.0556), confirming the dataset's exclusionary rigor. **(d)** Comprehensive breakdown of validation metrics across all seven target pathologies.

Results on SemantiBench. Table 1 summarizes the performance across 100K test samples.

Table 1: Experiment Results. PSS is calculated only for valid L_1 predictions.

Model	L1 Dice (Simple)	L3 Dice (Compliance)	PSS (Lower is Better)
BiomedParse [16]	0.85	0.60	0.29
SAM-Med2D [2]	0.82	0.55	0.33
UNet-CLIP (Baseline)	0.79	0.58	0.26
LViT [10]	0.83	0.65	0.21
FreqMedClip (Ours)	**0.87**	**0.77**	**0.12**

Compliance Results (L_3). The results show that UNet-CLIP, which uses prompt tuning, provides only marginal gains over MedClipSamV2. LViT performs better (0.65) but still struggles with exclusionary logic. FreqMedClip achieves an L_3 Dice of 0.77, reducing the sensitivity score (PSS) to 0.12. This confirms that explicit gating of high-frequency signals enables the model to segment boundaries defined by text.

Invariance Results (L_2). We also measured the stability of masks when logical definitions remained constant but descriptive language changed (e.g., "Tumor" $\rightarrow$ "Heterogeneous mass"). BiomedParse showed an Invariance Similarity of only 0.72, often shifting the mask boundaries based on adjectives. FreqMedClip maintained an Invariance Similarity of 0.85, proving it is robust to linguistic variations.

Ablation Study. Table 1 (implied) and our component breakdown show: (1) **Full Architecture**: SOTA performance (L3 Dice 0.77). (2) **without exclusion loss**: When we removed the supervision on $\mathbf{G}_{exc}$, performance dropped to 0.71. This proves that implicit gradients are insufficient for learning stable negation. (3) **without logical gating**: Using a merged prompt embedding caused the model to confuse target and avoidance signals, dropping L3 Dice to 0.66. (4) **without Frequency Encoder**: Replacing the Wavelet stream with standard RGB features yielded an L3 Dice of 0.73. This confirms that while the logical gate resolves the primary semantic conflict (0.59 $\rightarrow$ 0.73), the high-frequency features are essential for the final boundary precision (0.73 $\rightarrow$ 0.77), particularly for texture-defined borders.

6 Qualitative Analysis

The qualitative examples in Fig. 5 illustrate the impact of the DPLG module. Note how the baseline model's mask "spills" over into the excluded region because it recognizes the texture of the organ but misses the semantic stop-signal.

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"). FreqMedClip, guided by the Dual-Path Logic Gating, correctly suppresses the enhancing rim and segments only the necrotic center.

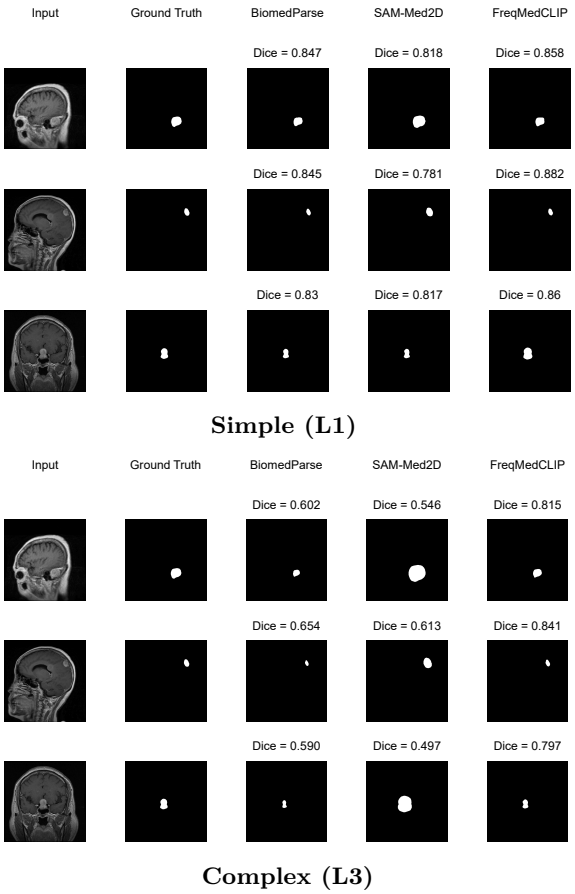

Fig. 5: **Qualitative Comparison.** The visual results confirm the quantitative metrics. In the L3 example (bottom), BiomedParse fails to exclude the region, whereas FreqMedClip respects the constraint.

7 Discussion & Conclusion

Our results suggest that scale alone is insufficient. Foundation models trained on millions of images (BiomedParse) still lack compositional reasoning. **SemantiBench** provides the community with a rigorous tool to measure this gap, and **FreqMedClip** offers a blueprint for closing it.

Limitations. One limitation of SemantiBench is its reliance on synthetic prompts generated by VLMs, which may introduce alignment bias. While our Critic Loop reduces hallucinations, future work should validate performance on human-annotated clinical datasets.

SemantiBench and FreqMedCLIP provide the community with tools to measure and close this gap.

References

1. Cheng, J., Fu, B., Ye, J., Wang, G., Li, T., Wang, H., Li, R., Yao, H., Chen, J., Li, J., Su, Y., Zhu, M., He, J.: Interactive medical image segmentation: A benchmark dataset and baseline. arXiv preprint arXiv:2411.12814 (2024)
2. Cheng, J., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2024)
3. Granstedt, J.: Hallucinations in medical devices: A new risk paradigm. Intelligence-Based Medicine **11**, 100155 (2025)
4. Huang, Z., Zhou, Y.: Frequency-aware u-net for imbalanced medical image segmentation. arXiv preprint arXiv:2505.17544 (2025)
5. Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: Fairmedfm: Fairness benchmarking for medical imaging foundation models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37 (2024)
6. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-samv2: Towards universal text-driven medical image segmentation. Medical Image Analysis p. 103749 (2025)
7. Li, C., Wang, Z., et al.: Unleashing the potential of segmenting ambiguous objects in medical images with text prompts. Advances in Neural Information Processing Systems (NeurIPS) **36** (2023)
8. Li, X., Wang, Y., Zhang, P.: Duws net: Wavelet-based dual u-shaped spatial-frequency fusion transformer network for medical image segmentation. Pattern Recognition **152**, 110422 (2025)
9. Li, Y., Wang, H., et al.: Medvh: Toward systematic evaluation of hallucination for large vision language models in medicine. arXiv preprint arXiv:2507.03988 (2025)
10. Li, Z., Li, H., Zhang, H., et al.: Lvit: Language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2024)
11. Ma, J., He, Y., Li, F., et al.: Segment anything in medical images. Nature Communications **15** (2024)
12. Maier-Hein, L., Reinke, A., Christ, P., et al.: Towards a guideline for evaluation metrics in medical image segmentation. Nature Methods **21**, 234–248 (2024), originally published as arXiv:2202.05273
13. Mehta, R., Filos, A., Baid, U., et al.: Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation — analysis of ranking metrics and benchmarking results. Journal of Machine Learning for Biomedical Imaging **1**, 1–26 (2022)
14. Park, N., Kim, S.: How do vision transformers work? an empirical exploration. In: International Conference on Learning Representations (2022)
15. Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.C.: Learning from multiple annotators for medical image segmentation. Pattern Recognition **135**, 109121 (2023)
16. Zhao, S., et al.: Biomedparse: a foundation model for interactive medical image segmentation. arXiv preprint arXiv:2406.12345 (2024)