MedGround: Bridging the Evidence Gap in Medical Vision–Language Models with Verified Grounding Data

Anonymous ACL submission

Abstract

Vision-**L**anguage **M**odels (VLMs) can generate convincing clinical narratives, yet frequently struggle to visually ground their statements. We posit this limitation arises from the scarcity of high-quality, large-scale clinical referring-localization pairs. To address this, we introduce MedGround, an automated pipeline that transforms segmentation resources into high-quality medical referring grounding data. Leveraging expert masks as spatial anchors, MedGround precisely derives localization targets, extracts shape and spatial cues, and guides VLMs to synthesize natural, clinically grounded queries that reflect morphology and location. To ensure data rigor, a multi-stage verification system integrates strict formatting checks, geometry- and medical-prior rules, and image-based visual judging to filter out ambiguous or visually unsupported samples. Finally, we present MedGround-35K, a novel multimodal medical dataset. Extensive experiments demonstrate that VLMs trained with MedGround-35K consistently achieve improved referring grounding performance, enhance multi-object semantic disambiguation, and exhibit strong generalization to unseen grounding settings. This work highlights MedGround as a scalable, data-driven approach to anchor medical language to verifiable visual evidence.[1]

1 Introduction

Recent VLMs have shown impressive capability in medical image understanding tasks such as report generation and clinical question answering (Li et al., 2023; Manzari et al., 2023; Nath et al., 2025; Sellergren et al., 2025; Liu et al., 2025; Yang et al., 2024; Tarhini et al., 2025; Wu et al., 2025). However, their linguistic fluency often outpaces fine-grained visual localization: a model may describe plausible findings while failing to identify where those findings appear in the image. Such visually unfaithful outputs undermine interpretability and can lead to "right-for-the-wrong-reason" predictions (Kim et al., 2025; Xie et al., 2023; Mahmood et al., 2025; Ostmeier et al., 2024; Pal et al., 2023; Huang et al., 2025; Xu, 2025). We characterize this mismatch as a cognitive–perceptual gap, where a model's cognitive competence is not mirrored by its perceptual grounding.

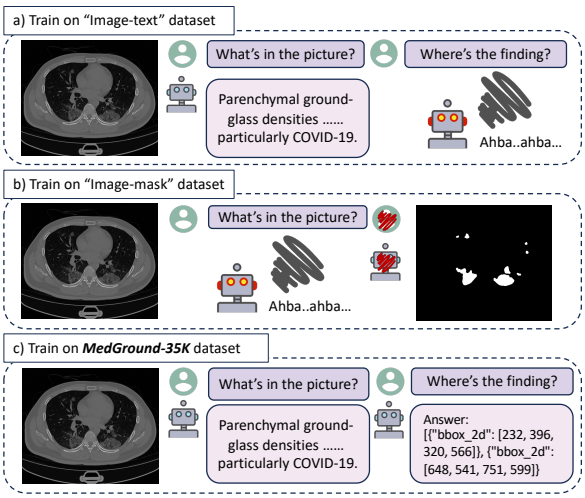

Figure 1: **Motivation of MedGround.** (a) Models trained on image-text pairs fail to *"speak with substance"* due to lack of grounding. (b) Segmentation-only training fails to achieve semantic understanding. (c) **MedGround** (Image-text-box triplets) activates the full potential of medical VLMs by bridging semantics and localization.

We argue that a key driver of this gap is the current data scarce. Medical data is plentiful in two largely disconnected forms: (1) image–text pairs from radiology reports, providing rich global semantics but weak spatial supervision; and (2) image–mask segmentation datasets, providing precise region annotations but impoverished language, often limited to a coarse category label. In contrast,

[1]Dataset and code will be released publicly upon acceptance.

large-scale datasets that pair natural referring expressions with explicit localization targets (boxes or masks) are scarce(Bannur et al., 2024; Li et al., 2024; Chen et al., 2025; Deng et al., 2025) .This scarcity limits the ability of VLMs to learn the alignment between clinically meaningful phrases and localized evidence.

However, it is challenging to construct medical referring grounding data on a scale. One seemingly naive approach is to ground existing reports by asking VLMs to localize findings. Yet current VLMs are often vision-weak at fine-grained grounding, so this direction risks inheriting and amplifying their spatial biases(Huang et al., 2024; Ge et al., 2025; Strudel et al., 2022; Bai et al., 2024; Zhang et al., 2025b). Instead, we reverse the process: starting from precise expert annotations in segmentation datasets, we use them as deterministic spatial anchors to synthesize referring expressions. This leverages the asymmetry of current models—strong language generation but weaker grounding—while maintaining spatial faithfulness through verification.

We introduce MedGround, a mask-guided semantic synthesis and verification pipeline that converts segmentation annotations into high-quality image–text–box triplets(MedGround-35K dataset) for medical referring expression grounding. MedGround effectively generates these triplets by leveraging expert masks to derive precise bounding boxes and extract geometric and spatial attributes. It then prompts a VLM to synthesize medically meaningful referring queries, followed by a multi-stage filtering process, including image-based VLM judging, to ensure data quality. Extensive experiments across multiple settings show that training with our constructed MedGround-35K dataset effectively improves medical referring grounding, and that the introduced clinically grounded semantics help VLMs better follow morphology- and location-aware descriptions for more reliable target disambiguation. These findings support MedGround as an effective and scalable supervision pipline for bridging the cognitive–perceptual gap.

In summary, our key contributions are:

- We propose MedGround, a scalable pipeline for synthesizing and verifying medically grounded referring queries anchored to expert annotations.
- We release MedGround-35K, covering eight datasets and multiple modalities, and demonstrate improvements in referring grounding, semantic disambiguation, and zero-shot transfer.
- We extensively tested various models using the MedGround-35K dataset and discovered that existing VLMs commonly struggle with fine-grained medical referring grounding tasks. However, training with our data significantly mitigates these issues.

2 Related Work

Our work relates to medical VLMs, referring grounding, and VLM-based synthesis and verifocation, so we review them below.

2.1 Medical Vision–Language Models

Medical VLMs have rapidly advanced with large-scale pretraining on radiology report corpora and biomedical image–text resources, followed by instruction tuning for clinical question answering and report generation(Li et al., 2023; Moor et al., 2023; Zhang et al., 2023). These models can produce fluent, clinically plausible narratives, but their outputs are not always evidence-aligned: models may describe findings without reliably localizing the corresponding visual regions, which limits interpretability and can lead to visually unfaithful reasoning(Pellegrini et al., 2023; Moll et al., 2025; Liu et al., 2023; Gundersen et al., 2025). This has motivated growing interest in grounding-aware evaluation and training signals that connect medical language to spatial evidence. In this work, we target this limitation by providing explicit referring-style localization supervision derived from expert segmentation annotations, enabling medical VLMs to better align morphology- and location-bearing phrases with concrete visual evidence.

2.2 Referring Expression Grounding

Referring expression grounding localizes an entity specified by a natural-language expression, emphasizing attribute understanding and spatial disambiguation. In natural images, large-scale referring and phrase grounding benchmarks have driven progress in models that bind text tokens to localized regions and handle fine-grained relational language(Liu et al., 2024; Li et al., 2022; Kamath et al., 2021). In medical imaging, however, referring expression grounding datasets remain scarce. Available supervision is typically split between image–text pairs (rich clinical narratives but weak spatial

Table 1: **Comparison of medical datasets.** ✓: provided; ✗: not provided; △: limited/weak or implicit/derived. Auto Anno. means the samples are collected autonomously.

Dataset	Modality	Text Type	Loc. Type	Granularity	Auto Anno.	#Anno.	Task
MS-CXR(Boecking et al., 2022)	CXR	✓ report phrases	✓ box	finding	✗	1,162	phrase grounding
Chest ImaGenome (Wu et al., 2021)	CXR	✓ structured phrases	△ region-aligned	region/finding	✗	1,256	region-text alignment
ChestX-ray8 (Wang et al., 2017)	CXR	△labels	✓box	disease	✗	888	disease localization
DeepLesion (Yan et al., 2018)	CT	✗ metadata	✓ box	lesion	✗	32,120	lesion detection
LIDC-IDRI (Armato et al., 2011)	CT	✗ attributes	✓ mask/contour	lesion	✗	875	nodule analysis
MedGround-35K (ours)	multi	✓ referring queries	✓ box	lesion/structure	✓	35,324	referring grounding

Table 2: **The statistics of the MedGround-35K.** The Anno. Tokens and Avg. Words columns show the total number of tokens and the average number of words for the medical grounding annotations regardless of task templates. The Modalities/Sources column shows the number of unique medical imaging sources/modalities involved in each split.

Split	#Images	Anno. Tokens	Avg. Words	Modality Ratio
Train	25.4k	2773.5k	12.0	Bacteria: 1.3%, CT: 8.2%, Dermoscopy: 9.9%, Nuclei: 20.5%, Ultrasound: 60.2%
Test	10.1k	1115.9k	12.7	Bacteria: 1.2%, CT: 18.3%, Dermoscopy: 11.8%, Nuclei: 25.3%, Ultrasound: 43.4%

alignment) and segmentation masks (precise localization but little to no language beyond class labels). This gap prevents medical VLMs from learning clinically meaningful referring cues such as morphology, laterality, anatomical sub-location, and multi-finding disambiguation(Bai et al., 2024; Deng et al., 2025; Liu et al., 2023; Zhang et al., 2025a; Yang et al., 2025). MedGround bridges this gap by converting segmentation annotations into scalable image–text–box supervision, explicitly training the model to follow clinically grounded referring descriptions.

2.3 VLMs-based Synthesis and Verification

VLMs-driven dataset construction has become a practical route to scale instruction and annotation resources, often paired with automated filtering, self-checking, or model-based judging to control noise. In medical settings, such synthesis is particularly fragile: hallucinated attributes, incorrect spatial relations, or visually unsupported statements can easily slip into training data and degrade evidence faithfulness(Pal et al., 2023; Yu et al., 2023; Khanna et al., 2023; Dong et al., 2023; Croxford et al., 2025; Yue et al., 2025). MedGround adopts a conservative synthesis principle: rather than asking a model to discover or localize findings from free-form text, we start from expert masks as deterministic spatial anchors and only synthesize referring queries conditioned on mask-derived geometry and spatial cues. We then apply multi-stage verification—format constraints, rule-based geometric/medical priors, and image-based judging—to filter ambiguous or visually unsupported samples. This design explicitly reduces the risk of inheriting vision-side localization bias while retaining the scalability benefits of VLMs-based generation.

3 MedGround

This section introduces **MedGround**, a mask-guided synthesis-and-verification pipeline (Fig. 2) that automatically converts segmentation annotations into high-quality *medical referring box grounding* datasets. Starting from expert masks, MedGround derives candidate ground-truth boxes, prompts a VLM to generate clinically grounded referring queries for randomly selected targets, and applies multi-stage verification to filter ambiguous or visually unsupported samples.

3.1 Dataset Definition

We produce MedGround-35K with MedGround pipline. Each example of MedGround-35K is a triplet (I, Q, B), where I is a medical image, Q is a referring query that describes one or multiple target regions, and B is the corresponding ground-truth set of 2D bounding boxes. Each box is represented as $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ and normalized to a 1000×1000 coordinate grid for consistent formatting across datasets and resolutions.

3.2 From Masks to Ground-Truth Box Lists

MedGround-35K is built from eight public segmentation datasets(Tab.5), spanning dermatology, microscopy, and radiological imaging. Each dataset provides expert pixel masks M. For each connected component in M, we deterministically derive a tight bounding box $b = \mathrm{bbox}(M)$, and collect all boxes within the same image into a candidate list $\mathcal{B} = \{b_1, \ldots, b_n\}$. This box list serves two roles: it provides the query generator (VLMs) with a *candidate target pool*, i.e., explicit regions

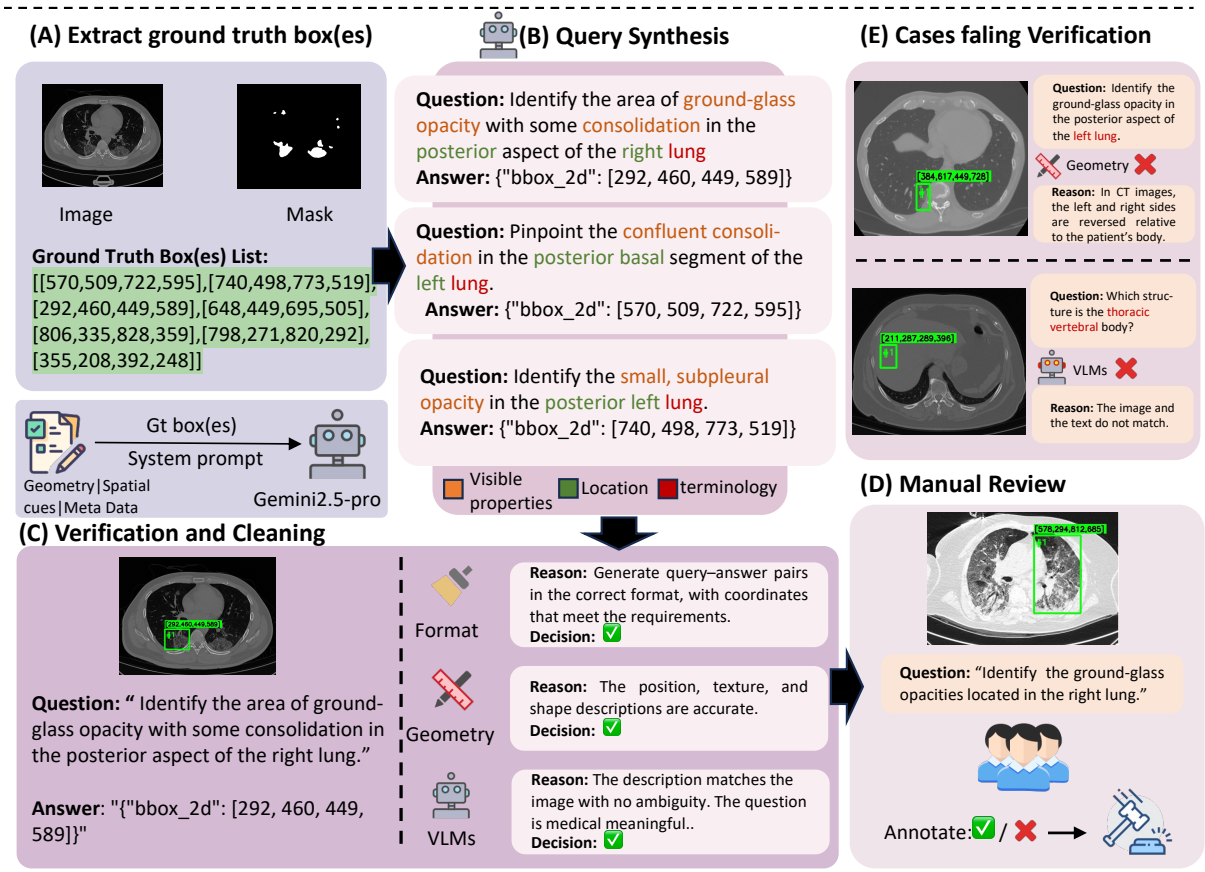

Figure 2: **MedGround pipeline.** (A) Convert segmentation masks into normalized ground-truth bounding box lists. (B) Use dataset-aware, mask-guided prompts to synthesize medically meaningful referring queries and select target box(es) as answers. (C) Perform multi-stage verification and cleaning (format/schema, geometry–location rules, and VLM-based grounding). (D) Conduct manual review for final quality control. (E) Cases falling verification.

to refer to when composing queries, from which one or multiple boxes are randomly selected as the intended target(s); and it serves as the *ground-truth localization anchor* during verification, supplying the reference boxes needed to judge whether the generated text is visually faithful and unambiguous with respect to the image.

3.3 Mask-Guided Query Synthesis

Given an image I and its box list $\mathcal{B}$, we first compute mask-derived attributes to constrain generation: **(1) geometry**: area ratio, width/height, aspect ratio, elongation/compactness proxies, **(2) spatial cues**: centroid, coarse bins such as left/right and upper/lower, optionally with dataset-specific anatomical conventions, and **(3) metadata**: imaging modality/domain and coarse category labels when available.

We then construct dataset-aware prompts and query VLMs to synthesize referring queries. The prompts explicitly condition on the image modality, and instruct the VLMs to produce questions that: (1) reflect *visible* properties (shape/texture/boundary appearance), (2) incorporate location cues when needed for disambiguation, and (3) use appropriate medical terminology while avoiding claims that cannot be justified from the image (e.g., etiology, pathology stage, or non-visible symptoms).

To encourage diversity and avoid trivial templates, we ask the VLMs to *randomly select* one or multiple target boxes from $\mathcal{B}$ and generate a corresponding query. Each generated sample is required to follow a fixed JSON schema containing the query text and the selected target box coordinates, enabling downstream automatic parsing.

3.4 Multi-Stage Data Verification and Cleaning

Since free-form generation may introduce ambiguity or visually unsupported descriptions, MedGround applies a multi-stage verifier to filter noisy samples.

Stage I: format and schema checks. We reject

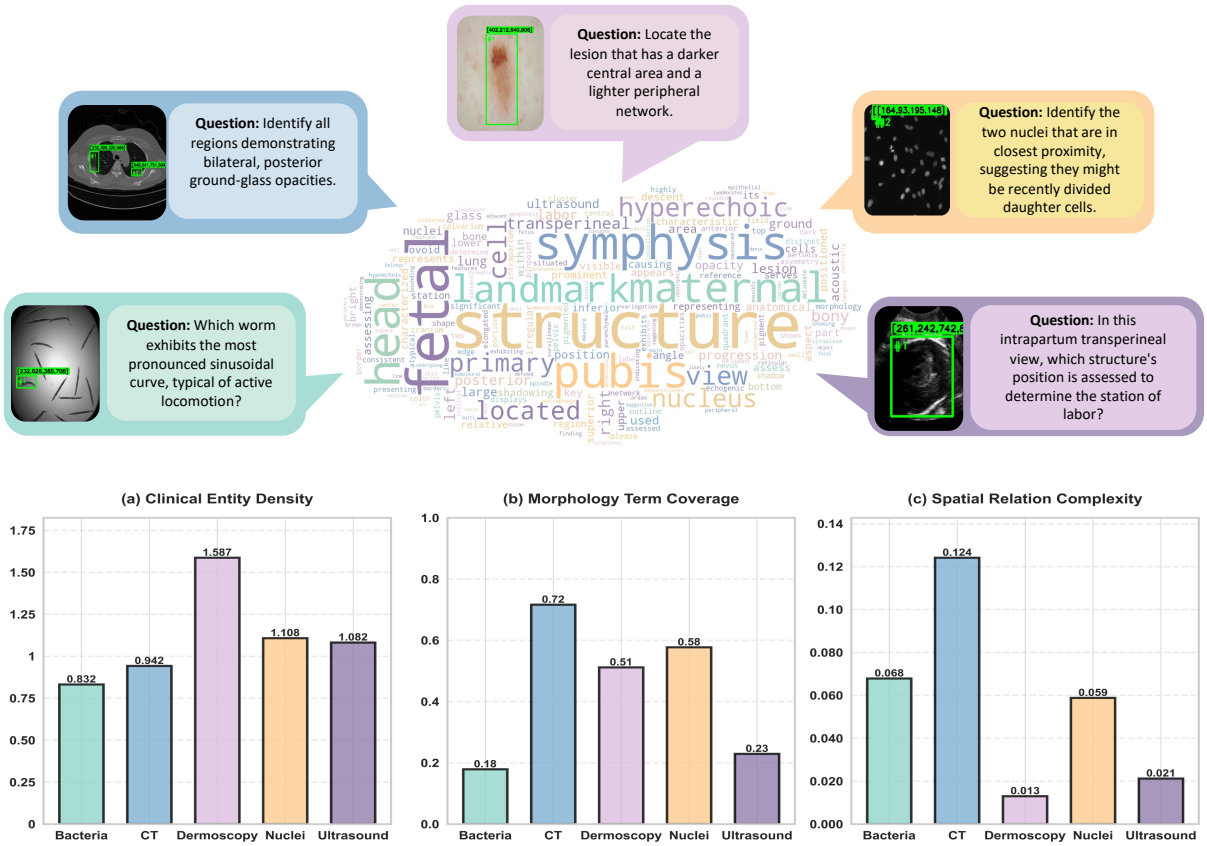

Figure 3: **Diversity and Linguistic Complexity of the MedGround dataset.** Up: The word cloud illustrates the distribution of medical terminology, anatomical landmarks, and clinical descriptors within the grounding annotations. Down: Comparative analysis of annotation richness across five distinct modalities based on three key metrics: (a) Clinical Entity Density, (b) Morphology Term Coverage, and (c) Spatial Relation Complexity.

generations that do not conform to the required JSON format, contain missing fields, or output invalid box indices/coordinates.

Stage II: rule-based validation with geometry and location constraints. We enforce consistency between textual descriptors and mask-derived attributes. For example, size adjectives must match area buckets, spatial phrases must match centroid bins, and prompts are constrained by domain-specific keyword allow/deny lists to prevent cross-domain leakage (e.g., thoracic anatomy terms in dermoscopy). Samples violating these rules are discarded.

Stage III: VLM-based consistency filtering. We performs image-based data cleaning using a VLMs as a semantic verifier. Recall that during synthesis, the generator first selects one or multiple target box(es) from the candidate pool $\mathcal{B}$ and then writes a question to refer to the selected region(s); therefore, the "answer" A of each QA pair is exactly the selected ground-truth box(es).

We feed the verifier the image together with the selected answer box(es) as an explicit localization anchor, and ask whether the visual content inside the box supports the question description. Concretely, the verifier is instructed to restate the observable attributes of the highlighted region and output a binary decision on whether the question is *correctly grounded* and *unambiguous*. If the verifier cannot recover the described cues from the given box, we treat the sample as hallucinated or ambiguous and discard it.

3.5 MedGround-35K

MedGround-35K comprises $35,480$ image–text–box triplets, with $25,420$ for training and $10,060$ for testing.

To characterize the semantic richness of the synthesized queries, we quantify three linguistically grounded properties: (1) **Clinical entity density:** Measured by extracting unique UMLS(Bodenreider, 2004) concepts via SciSpacy and normalizing by query length. (2) **Morphology coverage:** Assessed using a modality-aware

lexicon of appearance descriptors across CT, dermoscopy, microscopy, and ultrasound. (3) **Spatial complexity:** Calculated as the average frequency of spatial prepositions and relational phrases per query. Details semantic index of MedGround-35K are shown in Fig.3. In addition, we log the outcome of each verification stage and report the pass rate per stage in appendix(Tab.6).

3.6 Human Audit

To estimate the faithfulness of retained triplets and quantify residual noise, we conduct a full human audit on the entire MedGround-35K test set. Each triplet is independently reviewed by **three trained professional medical annotators** and marked as *good* if the referring query is clinically sensible and the target localization is visually consistent. We use **majority vote** to define high-confidence samples: a triplet is considered accepted if it receives at least two *good* votes ($\geq$2/3).

Overall, we audit 10,060 test triplets, and the overall majority-vote accept rate is **78%**, indicating that most synthesized queries are faithful and visually grounded. We provide the per-dataset audit breakdown in Appendix Tab. 7.

During human auditing, most rejected samples were due to ambiguity. We treat such ambiguity as a form of **natural noise** that is difficult to eliminate in real-world medical grounding. Moreover, these lower-agreement subsets constitute only a small fraction of MedGround. Therefore, we keep them in the dataset to preserve coverage and realism, while the appendix transparently documents their audit outcomes and typical error patterns.

Note: In the following experiments, we train models on the **original training split** to reflect realistic data conditions and avoid introducing human-selection bias. For evaluation, to ensure reliable and accurate measurement, we report results on the **human-verified test split**, where only triplets that pass the audit are retained.

4 Experiments

In this section, we evaluate the effectiveness of **MedGround-35K** in enhancing the fine-grained grounding capabilities of VLMs. We focus on whether the synthesized image–text–box triplets can empower models with evidence-grounded reasoning and improve medical semantic alignment beyond generic label-based supervision.

4.1 Experiment Settings

We evaluate models on three benchmarks that cover complementary aspects of medical grounding and generalization:

Medical Referring Grounding We use the full test split of MedGround-35K to assess medical referring grounding performance. Each sample consists of an image, a referring query, and the corresponding target box, requiring the model to localize the region described by the query.

Semantic Alignment To probe finer-grained semantic alignment, we construct a semantic test set from two datasets with multiple targets per image, namely MosMedPlus(Morozov et al., 2020) and FHPsAOP(Lu et al., 2022; Jieyun and Zhan-Hong, 2024). We select multi-target images and collect question pairs that refer to different targets within the same image. This setting explicitly tests whether a model can distinguish subtle semantic cues and ground them to the correct instance among several plausible regions.

Zero-shot Generalization We further evaluate zero-shot medical generalization on the QaTa-COV19(Degerli et al., 2021) dataset, where models are tested without task-specific tuning on it. This benchmark reflects out-of-distribution transfer and measures whether improvements from MedGround training generalize beyond the constructed data distribution.

Evaluation Metrics We utilize IoU to measure the precision of referring expression grounding.

Training Details We select MedGemma-27B, MedGemma-4B(Sellergren et al., 2025), Qwen2.5-VL-7B(Bai et al., 2025b), and Qwen3-VL-8B(Bai et al., 2025a) as the base models and fine-tune them on the MedGround-35K training data to investigate how fine-grained clinical semantics enhance their medical referring grounding capabilities. To rigorously isolate the impact of linguistic granularity, we also fine-tune these models on a label-based baseline where detailed referring expressions are replaced by coarse category names. All models are fine-tuned using LoRA(Hu et al., 2022) on 4 H20 GPUs for three epochs. (Further implementation details and hyper-parameters are provided in Appendix Tab.9$\sim$12).

VLMs Evaluated We compare with both general purpose VLMs and medical expert VLMs. During the evaluation, we manually craft grounding prompts suitable for these VLMs.

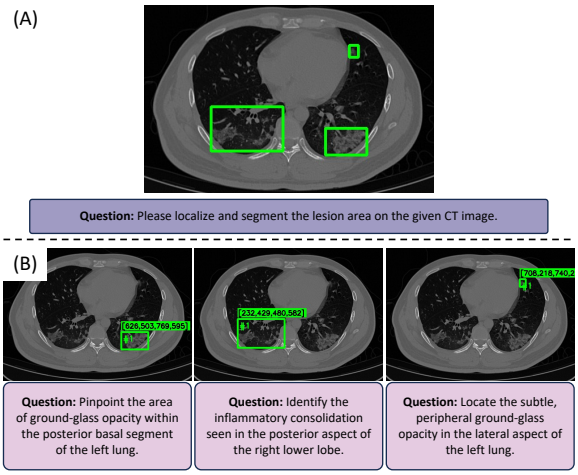

Figure 4: **Label-based Dataset vs. MedGround-35K.** (A) Label-based Dataset: conventional datasets typically group all ground-truth boxes under a single generic category label (e.g., "lesion") for the entire image. (B) MedGround-35K: provides distinct, fine-grained descriptive expressions for each localized box, capturing specific clinical nuances for individual regions.

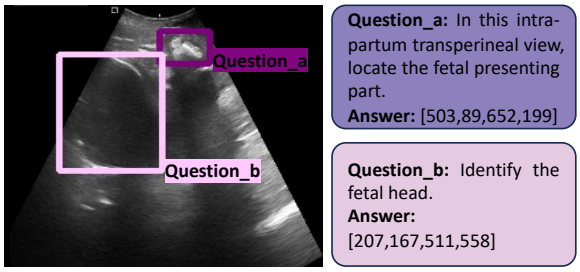

Figure 5: **Examples of the Semantic Sensitivity Testing Dataset**

4.2 Results and Analysis

MedGround effectively enhances VLMs' medical referring grounding capability. We evaluate the performance of VLM models trained on MedGround-35K. As shown in Tab.3, models fine-tuned with our knowledge-aware triplets demonstrate significant performance gains.

The experimental results demonstrate that fine-tuning on the MedGround-35K serves as a transformative catalyst for the medical referring grounding capabilities of both general and medical-specific VLMs. By providing high-quality spatial-instructional alignment, MedGround enables models to overcome the limitations of zero-shot reasoning. This consistent improvement across diverse modalities and model scales underscores MedGround's efficacy in distilling specialized medical spatial knowledge into large-scale models, effectively bridging the gap between general visual perception and precise clinical localization.

While MedGround-35K significantly improved overall performance, the MedGemma series showed minor regressions on DeepBACS(Spahn et al., 2022; Spahn and Heilemann, 2021) (e.g., -0.5 for MedGemma-27B). This is likely due to the domain shift between specialized fluorescence microscopy and the macro-level clinical imagery in MedGround. Fine-tuning may have caused a slight "feature forgetting" of niche microscopic traits while prioritizing general anatomical logic—a minor trade-off compared to the substantial gains achieved across the broader clinical spectrum.

MedGround injects fine-grained medical semantic knowledge into VLMs. We evaluate whether MedGround injects finer-grained clinical semantics than coarse label-level supervision using a Semantic Alignment setting. Concretely, we fine-tune the same backbone with LoRA on (i) MedGround-35K (clinically detailed referring expressions) and (ii) a label-based baseline built by pairing each annotated region with its coarse category name (Fig. 4). We then compare the two models on the Semantic Alignment Evaluation benchmark.

We propose **Semantic Sensitivity** to measure whether a model can follow subtle semantic differences in multi-target images. Each test case contains one image with multiple targets and two queries referring to different objects (Fig. 5). For each query, we compute the IoU between the predicted box and its corresponding ground-truth box; a target is counted as correct if IoU exceeds a threshold τ. The test case is scored as 1 only if *both* queries are correctly localized (otherwise 0). SS is the average score over all test cases.

As shown in Tab. 4, models without medical referring grounding training (w/o SFT) exhibit very low Semantic Sensitivity, indicating limited ability to follow query-specific semantics and localize the correct targets in multi-object clinical images. After fine-tuning, MedGround-35K SFT consistently outperforms the label-based SFT baseline on both FHPsAOP and MosMedPlus. This advantage stems from MedGround's fine-grained clinical semantics, whereas label-based supervision relies on coarse category names that are often insufficient to distinguish co-existing targets. These findings confirm that fine-grained clinical semantics are essential to bridge the "cognitive–perceptual gap," successfully equipping VLMs with the ability to associate specific morphology- and location-aware descrip-

Table 3: **Medical referring grounding performance across benchmarks.** We compare base models and their MedGround-finetuned counterparts. Colored deltas indicate changes over the corresponding base model (gain: purple, drop: pink).

Type	Model	Size	ISIC2016	BBBC010	BriFiSeg	CellNuclei	DeepBACS	FHPsAOP	MoNuSAC	MosMedPlus
General	Qwen2.5-VL-7B	7B	9.9	0.3	1.1	0.5	0.0	0.1	2.3	1.3
	Qwen3-VL-8B	8B	81.7	44.7	13.2	15.9	9.0	21.0	13.6	11.1
Medical VLMs	Lingshu-7B	7B	54.8	6.1	5.0	3.2	2.7	15.4	6.0	6.6
	MedGemma-4B	4B	50.4	4.0	3.0	2.3	0.6	17.1	3.5	3.5
	MedGemma-27B	27B	54.5	8.6	2.9	3.2	0.9	22.0	4.4	3.0
Finetuned	Qwen2.5-VL-7B	7B	83.0 (+73.1)	31.0 (+30.7)	20.2 (+19.1)	10.3 (+9.8)	7.1 (+7.1)	77.1 (+77.0)	10.9 (+8.6)	30.1 (+28.8)
	Qwen3-VL-8B	8B	86.4 (+4.7)	46.5 (+1.8)	24.6 (+11.4)	34.7 (+18.8)	23.0 (+14.0)	81.0 (+60.0)	13.5 (-0.1)	30.1 (+19.0)
	Lingshu-7B	7B	84.2 (+29.4)	44.1 (+38.0)	19.6 (+14.6)	9.7 (+6.5)	8.1 (+5.4)	79.0 (+63.6)	10.9 (+4.9)	36.3 (+29.7)
	MedGemma-4B	4B	71.2 (+20.8)	16.3 (+12.3)	6.2 (+3.2)	4.7 (+2.4)	0.0 (-0.6)	72.8 (+55.7)	5.3 (+1.8)	33.1 (+29.6)
	MedGemma-27B	27B	81.2 (+26.7)	26.8 (+18.2)	11.6 (+8.7)	11.2 (+8.0)	0.4 (-0.5)	80.6 (+58.6)	9.0 (+4.6)	39.3 (+36.3)

Table 4: **Comparison of the SS metric.** Grounding performance is reported in Semantic Sensitivity (%). Comparison across zero-shot baselines, label-based SFT, and MedGround SFT.

Base Model	Training Data	FHPsAOP	MosMedPlus
Qwen3VL-8B	w/o SFT	21.5	5.9
	Label-based SFT	45.3	5.8
	MedGround-35K SFT	**53.5**	**18.0**
MedGemma-4B	w/o SFT	2.3	1.4
	Label-based SFT	21.3	4.4
	MedGround-35K SFT	**55.9**	**13.1**
MedGemma-27B	w/o SFT	12.7	0.0
	Label-based SFT	27.0	2.3
	MedGround-35K SFT	**74.3**	**22.5**

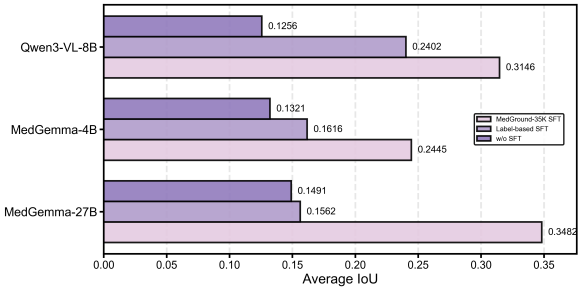

Figure 6: **Zero-shot Performance.** We evaluate different models on QaTa-COV19 dataset.

tions with their corresponding spatial anchors in multi-target environments.

MedGround Enables Zero-shot Generalization of VLMs. We further evaluate whether MedGround improves cross-dataset transfer by testing on an external unseen dataset, QaTa-COV19. We construct a fine-grained referring grounding test set from its referring statements and ground-truth boxes, and compare the performance of the MedGround-fine-tuned model against the base model on this benchmark.

The zero-shot evaluation on the external QaTa-COV19 dataset reveals that fine-tuning on MedGround-35K yields substantial and consistent improvements in cross-dataset transferability across all tested architectures. As illustrated in the Fig.6, the MedGround-35K enhanced models achieve remarkable performance gains compared to others. This significant performance leap on an entirely unseen dataset—achieved without any task-specific training—strongly demonstrates that MedGround does not merely facilitate dataset-specific memorization but rather imparts a robust, generalized medical spatial reasoning logic. Such results solidify MedGround's role as a critical foundational resource for enabling VLMs to generalize their referring grounding capabilities to novel clinical scenarios and emerging disease distributions.

5 Conclusion

We introduce **MedGround**, a mask-guided synthesis and verification pipeline that constructs medical referring grounding data as image–query–box triplets. By generating fine-grained referring expressions and filtering ambiguous or visually unsupported samples, MedGround provides semantically rich, visually verified supervision. Fine-tuning on MedGround-35K consistently improves grounding performance and transfer to unseen benchmarks, helping narrow the *cognitive–perceptual gap* by anchoring medical language to localized visual evidence. Since MedGround-35K is synthesized and verified from existing segmentation resources, it is highly scalable, enabling rapid extension to new modalities and anatomical systems. Beyond grounding, MedGround encourages models to justify clinical semantics with explicit spatial evidence, reducing reliance on linguistic priors and improving faithfulness.

Limitations

First, the VLM judge (Gemini-2.5-Pro) may introduce model-specific biases. Although the upstream stages of our pipeline are largely deterministic and rule-based, the final acceptance decision is made by a learned verifier and can reflect its preferences, failure modes, or sensitivity to prompt phrasing. Using alternative judges or ensembling multiple verifiers could further reduce this dependency, but is beyond the scope of this work. Second, our primary supervision is box-level rather than pixel-level. While bounding boxes provide scalable and broadly applicable grounding signals, they may not capture fine-grained boundaries or subtle morphology required by certain clinical applications (e.g., precise lesion extent or margin characterization). Third, synthesized queries may inherit LLM stylistic artifacts (e.g., phrasing patterns, verbosity, or implicit assumptions). We mitigate this with constrained prompting and verification, yet some residual biases may remain and could affect downstream generalization. Finally, MedGround is constructed from public segmentation datasets and therefore inherits their label spaces, imaging modalities, and population coverage; as a result, some anatomical regions, pathologies, and acquisition settings are underrepresented, and performance on out-of-distribution clinical data may be limited.

References

Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, and 1 others. 2011. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez Valle, and 1 others. 2022. Ms-cxr: Making the most of text semantics to improve biomedical vision-language processing. *URL: http://dx. doi. org/10.13026/b90j-vb87*.

Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, and 1 others. 2019. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. 2025. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24732–24741.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and 1 others. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*, pages 2025–04.

Aysen Degerli, Mete Ahishali, Mehmet Yamac, Serkan Kiranyaz, Muhammad EH Chowdhury, Khalid Hameed, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. 2021. Covid-19 infection map generation and detection from chest x-ray images. *Health information science and systems*, 9(1):15.

Ziye Deng, Ruihan He, Jiaxiang Liu, Yuan Wang, Zijie Meng, Songtao Jiang, Yong Xie, and Zuozhu Liu. 2025. Med-glip: Advancing medical language-image pre-training with large-scale grounded dataset. *arXiv preprint arXiv:2508.10528*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Hongyu Ge, Longkun Hao, Zihui Xu, Zhenxin Lin, Bin Li, Shoujun Zhou, Hongjin Zhao, and Yihang Liu. 2025. Clinkd: Cross-modal clinical knowledge distiller for multi-task medical images. *arXiv preprint arXiv:2502.05928*.

Benjamin Gundersen, Nicolas Deperrois, Samuel Ruiperez-Campillo, Thomas M Sutter, Julia E Vogt, Michael Moor, Farhad Nooralahzadeh, and Michael Krauthammer. 2025. Enhancing radiology report generation and visual grounding using reinforcement learning. *arXiv preprint arXiv:2512.10691*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. 2024. A refer-and-ground multimodal large language model for biomedicine. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 399–409. Springer.

Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. 2025. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3779–3787.

Bai Jieyun and Ou ZhanHong. 2024. Pubic symphysis-fetal head segmentation and angle of progression.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790.

Sameer Khanna, Adam Dejl, Kibo Yoon, Steven QH Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. 2023. Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In *Machine learning for healthcare conference*, pages 381–402. PMLR.

Soo Yong Kim, Suin Cho, Vincent-Daniel Yun, and Gyeongyeon Hwang. 2025. Medclm: Learning to localize and reason via a cot-curriculum in medical vision-language models. *arXiv preprint arXiv:2510.04477*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, and 1 others. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975.

Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. 2024. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–90. Springer.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, and 1 others. 2025. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. 2012. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637.

Yaosheng Lu, Mengqiang Zhou, Dengjiang Zhi, Minghong Zhou, Xiaosong Jiang, Ruiyu Qiu, Zhanhong Ou, Huijin Wang, Di Qiu, Mei Zhong, and 1 others. 2022. The jnu-ifm dataset for segmenting pubic symphysis-fetal head. *Data in brief*, 41:107904.

Razi Mahmood, Pingkun Yan, Diego Machado Reyes, Ge Wang, Mannudeep K Kalra, Parisa Kaviani, Joy T Wu, and Tanveer Syeda-Mahmood. 2025. Evaluating automated radiology report quality through fine-grained phrasal grounding of clinical findings. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. 2023. Medvit: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157:106791.

Gendarme Mathieu, El Debs Bachir, and 1 others. 2022. Brifiseg: a deep learning-based method for semantic and instance segmentation of nuclei in brightfield images. *arXiv preprint arXiv:2211.03072*.

Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazhen Pan, Daniel Rueckert, Lisa C Adams, and Keno K Bressem. 2025. Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. *arXiv preprint arXiv:2510.11196*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: a multimodal medical few-shot learner (2023). *URL: https://arxiv.org/abs/2307.15189*.

Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, and 1 others. 2025. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14788–14798.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain

hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Christoph Spahn, Estibaliz Gómez-de Mariscal, Romain F Laine, Pedro M Pereira, Lucas von Chamier, Mia Conduit, Mariana G Pinho, Guillaume Jacquemet, Séamus Holden, Mike Heilemann, and 1 others. 2022. Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches. *Communications Biology*, 5(1):688.

Christoph Spahn and Mike Heilemann. 2021. Deepbacs–escherichia coli bright field segmentation dataset.

Robin Strudel, Ivan Laptev, and Cordelia Schmid. 2022. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*.

Ali A Tarhini, Palak Dave, and Issam El Naqa. 2025. General artificial intelligence for the diagnosis and treatment of cancer: the rise of foundation models. *BJR| Artificial Intelligence*, 2(1):ubaf015.

Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, and 1 others. 2021. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, and 1 others. 2021. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: a systematic review with large language models and beyond. *MedRxiv*.

Jenny Xu. 2025. Uncertainty estimation in large vision language models for automated radiology report generation. In *Proceedings of the 4th Machine Learning for Health Symposium. PMLR*, pages 1039–1052.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. 2018. Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, and 1 others. 2024. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Xuzheng Yang, Junzhuo Liu, Peng Wang, Guoqing Wang, Yang Yang, and Heng Tao Shen. 2025. New dataset and methods for fine-grained compositional referring expression comprehension via specialist-mllm collaboration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Jingkun Yue, Siqi Zhang, Zinan Jia, Huihuan Xu, Zongbo Han, Xiaohong Liu, and Guangyu Wang. 2025. Medsg-bench: A benchmark for medical image sequences grounding. *arXiv preprint arXiv:2505.11852*.

Wenjun Zhang, Shekhar S Chandra, and Aaron Nicolson. 2025a. Anatomical grounding pre-training for medical phrase grounding. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Wenjun Zhang, Shekhar S Chandra, and Aaron Nicolson. 2025b. Generalized medical phrase grounding. *arXiv preprint arXiv:2512.01085*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

A Datasets

A.1 Data source

MedGround-35K is constructed by repurposing annotations from eight publicly available medical segmentation datasets, covering diverse imaging modalities and clinical scenarios, including BBBC010(Ljosa et al., 2012), BriFiSeg(Mathieu et al., 2022), CellNuclei(Caicedo et al., 2019), DeepBacs(Spahn et al., 2022), FHPsAOP(Lu et al., 2022), ISIC2016(Codella et al., 2018), MoNuSAC(Verma et al., 2021), and MosMedPlus(Morozov et al., 2020). For each source dataset, we use its original train/val/test split as the raw data pool before applying our MedGround construction and verification pipeline. Detailed split statistics of these sources are summarized in Tab. 5.

Table 5: Raw split sizes of segmentation data sources before MedGround construction.

Dataset	Train	Val	Test	Total
BBBC010	70	10	20	100
Brifiseg	38,463	4,244	576	43,283
CellNuclei	469	67	134	670
DeepBacs	17	2	15	34
DynamicNuclear	4,950	1,417	717	7,084
FHPsAOP	5,600	800	1,600	8,000
ISIC2016	810	90	379	1,279
MoNuSAC	359	35	249	643
MosMedPlus	1,910	271	547	2,457
Total	52,648	6,936	4,237	63,550

A.2 Pass rate of multi-stage verifaction pipline

We analyze rejection reasons across stages (I: format, II: rule-based checks, III: VLM judging) in Tab. 6. Stages I and II retain almost all samples across datasets ($\sim$ 92–100%), indicating that invalid formatting and obvious rule violations account for only a small fraction of failures. In contrast, Stage III is substantially more selective and becomes the main source of filtering: retention drops to 48–67% for several microscopy datasets (e.g., BBBC010, CellNuclei, DeepBACS, BriFiSeg), while remaining high for more visually distinctive settings such as FHPSAOP and ISIC2016 ($\sim$88–90%). Overall, the final retention is about 80% on both splits (80.9% train, 80.3% test), suggesting that the VLM judge primarily removes samples that are *ambiguous* (multiple plausible referents) or *visually unsupported* (the description does not clearly match the highlighted target). This aligns with our human audit, where most rejected cases are due to ambiguity, and supports that MedGround's verification prioritizes unambiguous, evidence-grounded referring expressions rather than merely enforcing formatting constraints.

A.3 Human Audit Results and Failure Analysis

From the overall statistics, the lower pass rates are mainly observed in the **nuclei segmentation** datasets. By analyzing the auditors' rejection reasons, we find that most failures stem from "ambiguous matching"—the referring description does not admit a unique target in the image, leading auditors to judge the image–text pair as not well aligned. This is consistent with the inherent properties of nuclei images, where scenes contain numerous visually similar instances and the language often relies on low-dimensional attributes that are insufficient for unique identification.

A.4 Train–Test Distribution Analysis in Three Semantic Dimensions

We quantify linguistic properties relevant to grounding(Fig. 7): Clinical entity density using medical entity recognition/linking tools (e.g., UMLS-based pipelines). Morphology term coverage using a curated lexicon of appearance descriptors. Spatial relation complexity via counts of spatial prepositions and relational phrases. These analyses show that MedGround queries contain substantially richer morphology and spatial language than category-only prompts, providing more informative supervision for grounding.

First, clinical entity density remains stable between splits for most datasets, indicating similar average numbers of clinically relevant referents per sample. This stability is important for evaluating grounding in multi-target scenes, since the number of co-existing entities directly affects ambiguity and search difficulty.

Second, morphology term coverage shows noticeable cross-dataset variation and a mild train–test gap for some sources. In particular, datasets such as MosMedPlus exhibit higher coverage in the test split, implying that the audited test set contains richer morphological descriptions and thus places greater emphasis on fine-grained visual-semantic alignment. Conversely, nuclei-related datasets (e.g., BBBC010, DeepBACS, MoNuSAC) generally present lower morphology coverage, re-

Table 6: **Pass rates of each verification stage across datasets.** We report the retention percentage and remaining sample count after Stages A/B/C for both the training and test splits.

Stage	BBBC010		BriFiSeg		CellNuclei		DeepBACS		FHPSAOP		ISIC2016		MoNuSAC		MosMedPlus		Total	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Stage I	97.2%	97.0%	98.3%	98.4%	98.1%	98.4%	98.2%	97.2%	100.0%	100.0%	92.1%	92.4%	96.2%	96.7%	96.4%	94.8%	98.3%	97.6%
Stage II	96.8%	97.0%	98.1%	98.0%	97.9%	98.3%	98.2%	97.2%	99.4%	99.2%	92.0%	92.3%	95.4%	95.8%	95.6%	94.5%	97.8%	97.1%
Stage III	56.9%	49.3%	67.1%	65.9%	57.1%	56.0%	48.2%	47.2%	89.7%	89.3%	88.2%	88.5%	63.8%	60.6%	78.4%	94.5%	80.9%	80.3%
#Remaining	265	66	2369	1212	1651	479	54	51	15291	4363	2523	1191	1193	855	2074	1843	25420	10060

Table 7: **Human audit on the MedGround test split.** Each sample is reviewed by three auditors. We report the distribution of *good* votes and the majority-pass rate (Good Ratio, $\geq$2/3 *good*).

Dataset	Total	3 good	2 good	1 good	0 good	Good Ratio
ISIC2016_512_test	1,141	938 (82.2%)	142 (12.4%)	34 (3.0%)	27 (2.4%)	94.65%
mosmedplus_512_test	1,843	1,531 (83.1%)	57 (3.1%)	49 (2.7%)	206 (11.2%)	86.16%
fhpsaop_256_test	4,363	3,167 (72.6%)	479 (11.0%)	233 (5.3%)	484 (11.1%)	83.61%
brifiseg_512_test	1,212	334 (27.6%)	500 (41.3%)	283 (23.3%)	95 (7.8%)	68.81%
bbbc010_512_test	66	34 (51.5%)	10 (15.2%)	13 (19.7%)	9 (13.6%)	66.67%
cellnuclei_256_test	479	88 (18.4%)	147 (30.7%)	158 (33.0%)	86 (18.0%)	49.06%
monusac_512_test	855	129 (15.1%)	267 (31.2%)	316 (37.0%)	143 (16.7%)	46.32%
deepbacs_512_test	51	8 (15.7%)	14 (27.5%)	20 (39.2%)	9 (17.6%)	43.14%

flecting the inherent limitation that many instances are visually similar and hard to differentiate with morphology alone.

Third, spatial relation complexity is consistently highest for MosMedPlus in both splits, reflecting that CT findings often require more location-aware language (e.g., lobes, peripheral vs. central, bilateral distribution) to uniquely specify targets. In contrast, pathology and dermoscopy datasets tend to have lower spatial complexity, consistent with their relatively localized lesions or simpler spatial context. This three-dimensional analysis highlights that different modalities stress different aspects of grounding: microscopy emphasizes dense-instance disambiguation, while CT emphasizes spatial reasoning, and the test split preserves (or slightly strengthens) these challenges.

B Implementation Details

B.1 Fine-Tuning Details

We select five open-source VLM backbones as our base models: Qwen2.5-VL(Bai et al., 2025b), Qwen3-VL(Bai et al., 2025a), MedGemma-4B, MedGemma-27B(Sellergren et al., 2025), and Lingshu-7B(Xu et al., 2025). We report the training hyperparameters for each model in Tab.8 $\sim$ Tab.12.

Our training data are constructed from multiple segmentation datasets spanning different medical domains and imaging modalities(Tab.5). As a result, the collected training corpus exhibits notable cross-domain imbalance (e.g., microscopy data are substantially larger than some radiology or pathology sources). In our experiments, we do not apply explicit re-balancing strategies (e.g., re-sampling or per-domain weighting); instead, we directly fine-tune models on the naturally imbalanced mixture to reflect realistic data availability.

All training samples are obtained solely through our automatic verification pipeline, without any additional human filtering. In contrast, for evaluation we use a test set where all samples are manually audited to ensure correctness and unambiguity. This protocol isolates the effect of MedGround data quality and avoids overestimating performance due to noisy automatic annotations.

C Samples

C.1 MedGround-35K Examples

Fig. 8 presents representative examples from MedGround-35K across imaging modalities. Each example consists of a medical image, a fine-grained referring expression, and the corresponding target annotation, illustrating the diversity of morphology- and location-aware clinical semantics in our dataset.

Table 8: Training hyper-parameters used for fine-tuning Lingshu-7B in our experiments.

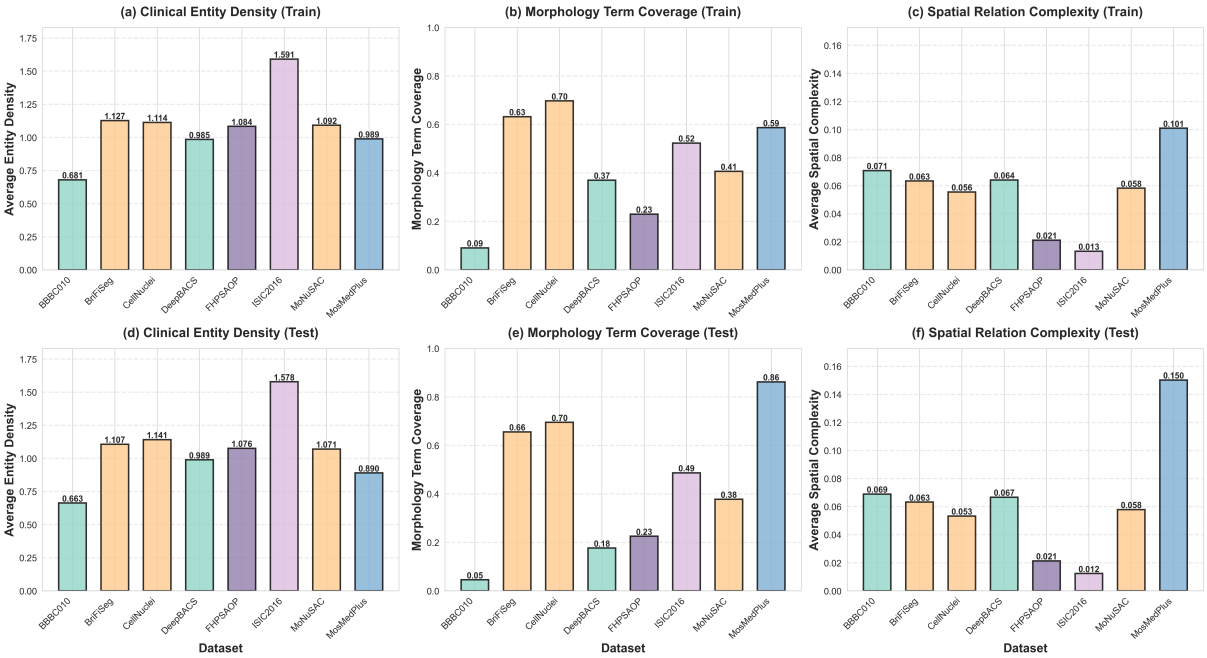

Figure 7: Three-dimensional analysis comparing the training split (top row, a–c) and test split (bottom row, d–f) across eight medical imaging datasets. Colors: blue=CT, dark purple=ultrasound, light purple=dermoscopy, green=nuclei, orange=bacteria.

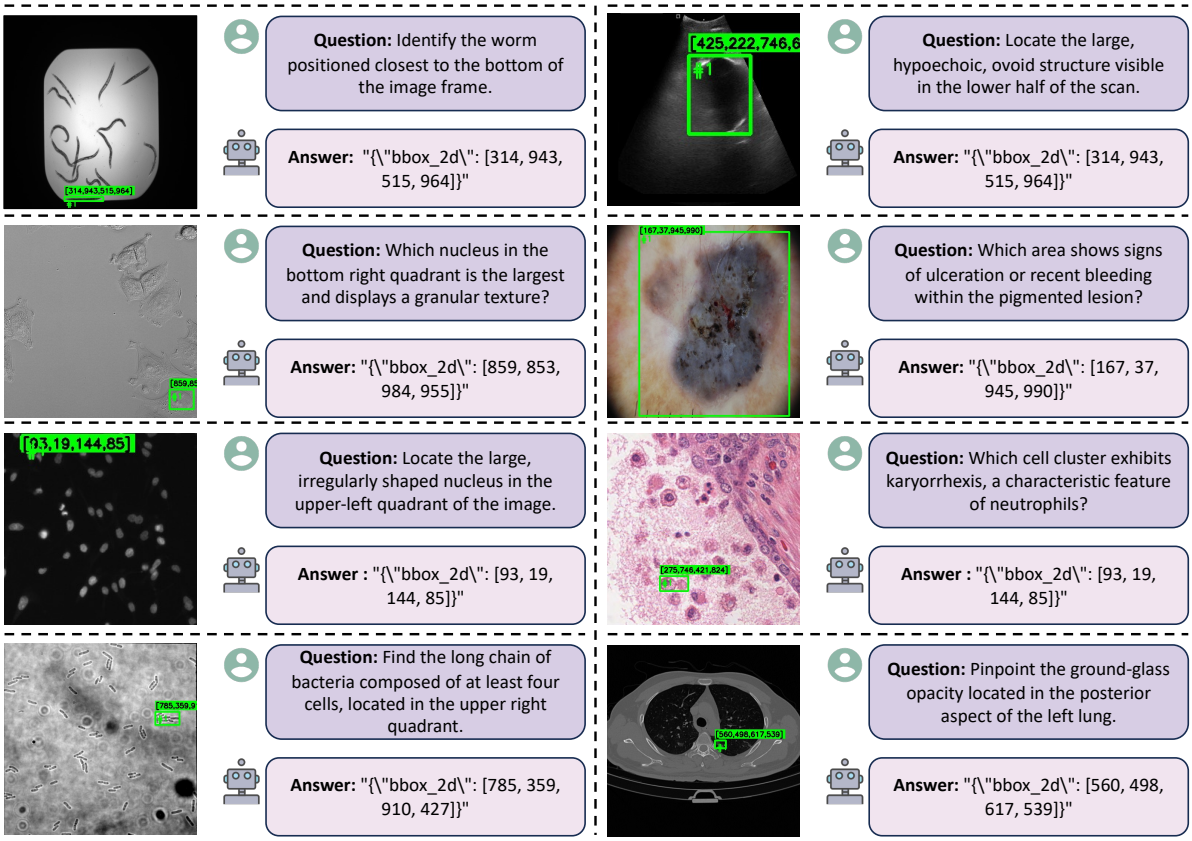

Figure 8: Examples of MedGround-35K.

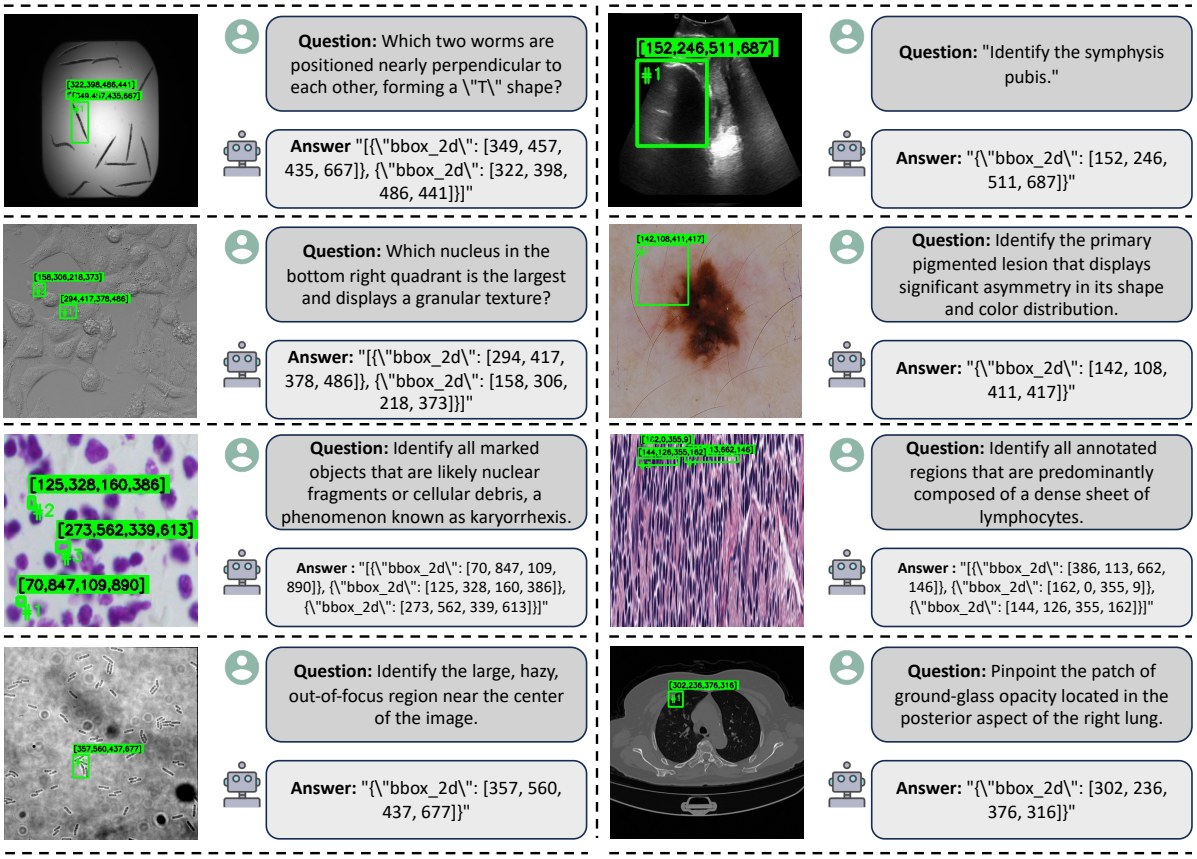

Figure 9: Examples of MedGround-35K.

Hyper-Parameter	Value
Backbone	Lingshu-7B
Training stage	SFT
Fine-tuning method	LoRA
LoRA Rank	8
LoRA Alpha	32
LoRA Target	all
Epoch	3
#GPUs	4
Per-device batch size	4
Gradient accumulation	4
Global batch size (effective)	64
Learning rate	2×10^{-4}
LR scheduler	Cosine
Warm-up ratio	0.1
Model max length	2048
Precision	BF16
Gradient checkpointing	Enabled
Random seed	42

Table 9: Training hyper-parameters used for fine-tuning Qwen2.5-VL-7B in our experiments.

Hyper-Parameter	Value
Backbone	Qwen2.5-VL-7B-Instruct
Training stage	SFT
Fine-tuning method	LoRA
LoRA Rank	8
LoRA Alpha	32
LoRA Target	all
Epoch	3
#GPUs	4
Per-device batch size	4
Gradient accumulation	4
Global batch size (effective)	64
Learning rate	2×10^{-4}
LR scheduler	Cosine
Warm-up ratio	0.1
Model max length	2048
Precision	BF16
Gradient checkpointing	Enabled
Random seed	42

Table 10: Training hyper-parameters used for fine-tuning MedGemma-4B in our experiments.

Hyper-Parameter	Value
Backbone	MedGemma-4B-IT
Training stage	SFT
Fine-tuning method	LoRA
LoRA Rank	8
LoRA Alpha	32
LoRA Target	all
Epoch	3
#GPUs	8
Per-device batch size	16
Gradient accumulation	1
Global batch size (effective)	128
Learning rate	2×10^{-4}
LR scheduler	Cosine
Warm-up ratio	0.1
Model max length	2048
Precision	BF16
Gradient checkpointing	Enabled
Random seed	42

Table 11: Training hyper-parameters used for fine-tuning MedGemma-27B in our experiments.

Hyper-Parameter	Value
Backbone	MedGemma-27B-IT
Training stage	SFT
Fine-tuning method	LoRA
LoRA Rank	8
LoRA Alpha	32
LoRA Target	all
Epoch	3
#GPUs	4
Per-device batch size	8
Gradient accumulation	2
Global batch size (effective)	64
Learning rate	2×10^{-4}
LR scheduler	Cosine
Warm-up ratio	0.1
Model max length	2048
Precision	BF16
Gradient checkpointing	Enabled
Random seed	42

Table 12: Training hyper-parameters used for fine-tuning Qwen3-VL-8B in our experiments.

Hyper-Parameter	Value
Backbone	Qwen3-VL-8B-Instruct
Training stage	SFT
Fine-tuning method	LoRA
LoRA Rank	8
LoRA Alpha	32
LoRA Target	all
Epoch	3
#GPUs	8
Per-device batch size	32
Gradient accumulation	1
Global batch size (effective)	256
Learning rate	2×10^{-4}
LR scheduler	Cosine
Warm-up ratio	0.1
Model max length	2048
Precision	BF16
Gradient checkpointing	Enabled
Random seed	42

C.2 Failure Cases in Manual Screening

The Fig.9 presents failure cases identified during manual screening/curation. These examples were flagged as incorrect due to issues such as question–image mismatch, mislocalized or improperly sized bounding boxes, incorrect object counting (marking multiple regions when a single target is required, or vice versa), and misinterpretation of domain-specific findings (e.g., confusing blur, noise, or debris with true anatomical/pathological structures). We include these samples to illustrate common error patterns and to motivate stricter quality control and refinement of our annotation guidelines.

D Details of General Generate Prompt

The Fig.10 shows the general system prompt we used to generate our synthetic data. This prompt serves as a unified instruction template that standardizes the model's role, output format, and task-specific constraints, ensuring consistency and controllability across different synthesis scenarios.

E Potential Use of MedGround

We further argue that referring grounding is intrinsically more effective at eliciting visual reasoning than conventional descriptive supervision. In descriptive tasks, models can often produce plausible responses by leaning on linguistic priors or memorized report patterns. In contrast, referring grounding requires the model to *commit* to a specific region, forcing clinical claims to be supported by explicit pixel-level evidence. This evidence-binding property makes grounding not only a practical capability for downstream localization, but also a training signal that can stimulate broader diagnostic reasoning by aligning clinical semantics with visual anchors. We therefore view MedGround as a general-purpose resource that may benefit a wide range of medical VLM applications, including faithful visual question answering, evidence-based reporting, and interactive decision support.

"""
###Role and Goal
You are a senior biomedical imaging analyst and an expert educator. Your role is an "Imaging Question Constructor".

The Task: Given a biomedical image, a detailed clinical description, and a list of Ground Truth (GT) bounding boxes, you must generate professional localization Q&A pairs. You are required to deconstruct the description, extract key medical attributes, and recombine them into new, targeted questions that map directly to the provided bounding boxes.

###Construction Strategy (Deconstruction & Synthesis)
- Deconstruct Description: Break down the provided description into atomic clinical elements: cell types, pathological features (e.g., morphology, staining), spatial arrangements, and diagnostic significance.
- Map to GT: Correlate specific segments of the description with the provided Bounding Box List.
- Synthesize Questions: Create new, professionally phrased questions that require identifying:
- Single Target: Isolate one specific entity based on a unique attribute mentioned in the description.
- Multiple Targets (Subsets): Group targets that share common features (e.g., "all cells with nuclear pleomorphism").
- Full Set: Identify all instances of a category described in the text.
- Recombine Context: Do not just repeat the description. Rephrase and combine attributes (e.g., if the description mentions "large cells" and "blue nuclei," ask to "Locate the macro-lymphocytes exhibiting intense basophilic nuclear staining").

###Strict Rules
- Data Integrity: The coordinates in the Answer: field MUST be exact copies from the provided Bounding Box List. Hallucination of coordinates is strictly forbidden.
- Medical Depth: Use professional terminology (e.g., "atypical mitotic figures," "stromal fibrosis," "vacuolated cytoplasm"). The questions should sound like they were written by a pathologist or radiologist.
- Logic-Description Mapping: Every question must be grounded in the facts provided in the description.
- Clean Output: Output ONLY Question/Answer pairs. No introductory text, no meta-explanation.
- The input you're given consists of CT images. Please carefully observe the cardiac silhouette's position. Note that generally, the right lung is on the left side of the image, and the left lung is on the right side of the image. The heart is on the person's left, which corresponds to the right side of the image. It is crucial not to confuse the left and right sides!

###Answer Format
- Single target: Answer: {"bbox_2d": [x1, y1, x2, y2]}
- Multiple targets: Answer: [{"bbox_2d": [x1, y1, x2, y2]}, {"bbox_2d": [x3, y3, x4, y4]}]

###Example Input
Image: [A pathology slide of lung tissue]
Description: bilateral pulmonary infection, two infected areas, all left lung and all right lung.
Bboxes:[[227,165,482,781],[535,89,883,638]]

Example Output
Question: Pinpoint the region within the thoracic cavity demonstrating infectious consolidation affecting the entirety of the left pulmonary parenchyma.
Answer: {"bbox_2d": [535, 89, 883, 638]}

Question: Identify the pulmonary segment exhibiting diffuse infectious changes within the right hemithorax.
Answer: {"bbox_2d": [227, 165, 482, 781]}

Question: Locate all areas indicative of bilateral infectious processes involving the pulmonary fields.
Answer: [{"bbox_2d": [227, 165, 482, 781]}, {"bbox_2d": [535, 89, 883, 638]}]

Figure 10: Detail of general generate system prompt.