Scene Grounding in Dense Visual Environments

1. Abstract

This report presents a CPU-only pipeline for localizing semantically relevant regions in dense scenes using natural-language queries. The system combines GroundingDINO for text-conditioned detection, optional CLIP-based re-ranking, sliding-window tiling with fusion for large images, and SAM for mask-level refinement. The goal is to robustly return a bounding box and a cropped region (and optionally a mask) that best matches the query.

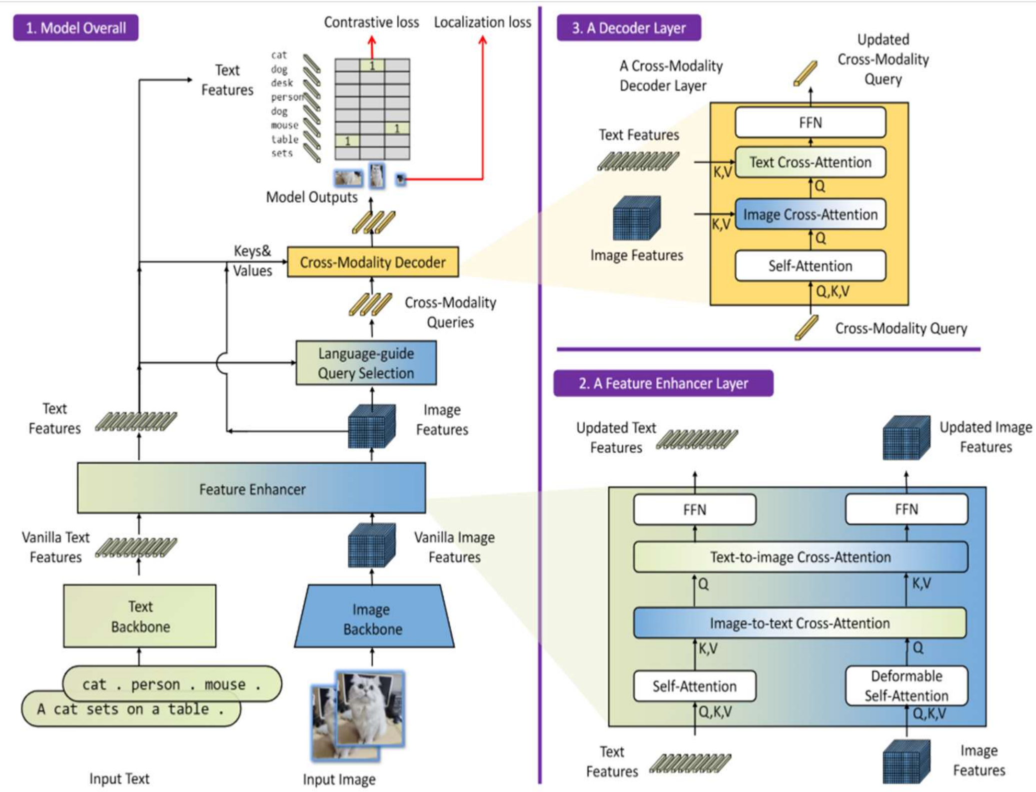


Figure 3. The framework of Grounding DINO. We present the overall framework, a feature enhancer layer, and a decoder layer in block 1, block 2, and block 3, respectively.

2. Introduction and Problem Statement

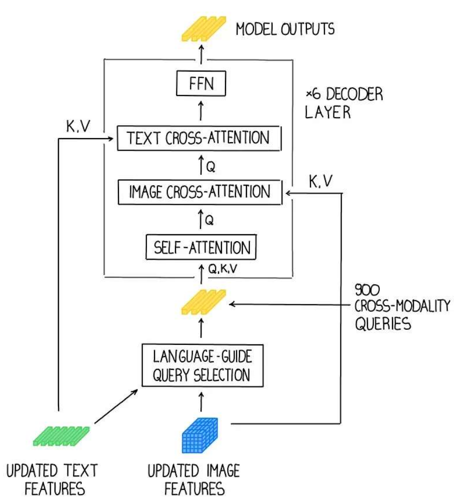
Dense scenes such as markets or stations contain overlapping objects and concurrent activities. Given an image and a free-form text description, the system outputs the region that best corresponds to the described interaction. CPU-only deployment eases portability and reproducibility.

Contributions:

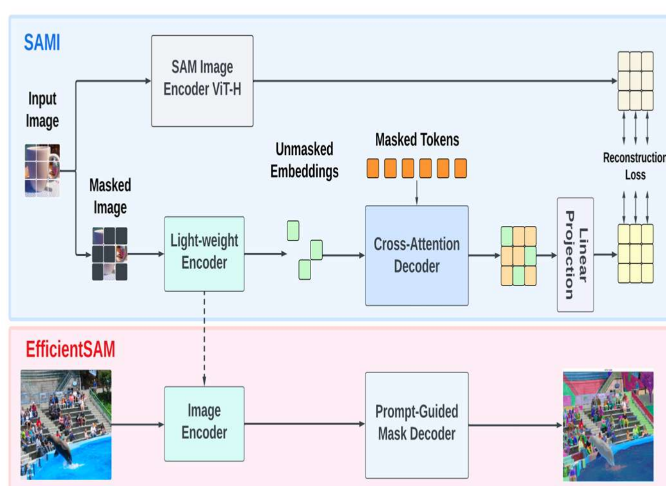
- Practical CPU pipeline combining detection, language grounding, and segmentation.
- Optional CLIP re-ranking and tiling with lightweight fusion (Soft-NMS, WBF).
- Reproducible setup with asset-fetcher script for weights and repos.

3. System Overview

1) GroundingDINO produces candidate boxes conditioned on the text prompt.
2) CLIP Re-ranking (optional) evaluates cropped proposals against the text embedding.
3) Tiling + Fusion (optional) improves coverage on large images; box candidates are fused.
4) SAM refines the final region by predicting a mask from the chosen bounding box.



GROUNDING-DINO



SAM & efficient-SAM

4. Methods

4.1 GroundingDINO (Text-Conditioned Detection):

- Input: RGB image, query text.
- Output: Bounding boxes with confidence scores.
- Apply confidence threshold, size filtering, and NMS (torchvision) or Soft-NMS.

4.2 CLIP Re-ranking (CPU):

- Crop each candidate box and encode with CLIP; encode the query text.
- Compute cosine similarity and fuse with detector confidence.

4.3 Tiling + Fusion:

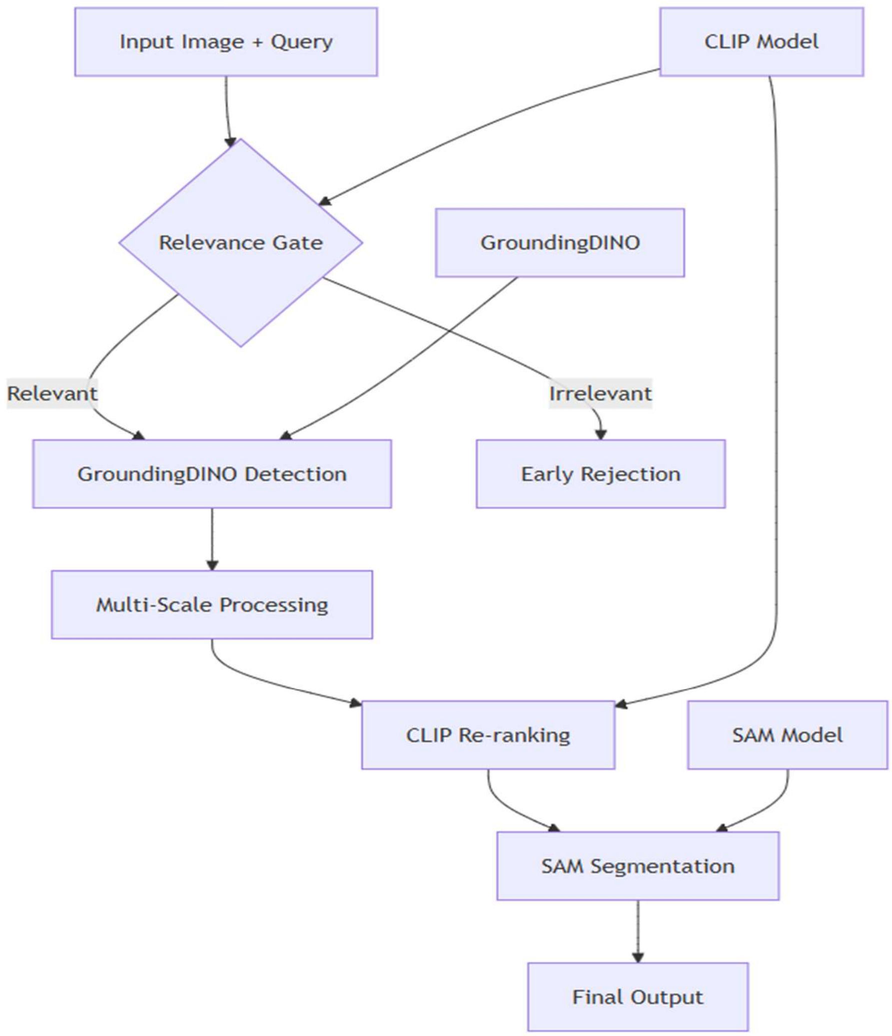
- Sliding-window tiles with overlap, detection per tile, and global coordinate mapping.
- Fuse boxes with Soft-NMS or WBF.

4.4 SAM Mask Refinement:

- Use selected box as a prompt to SAM ViT-H.
- Export binary mask, overlay, and RGBA cutout.

4.5 Negative Queries Handling:

- Support optional negative keywords with CLIP penalty or heuristic filter.



FLOWCHART/PIPELINE OF THE SOLUTION

5. Implementation Details

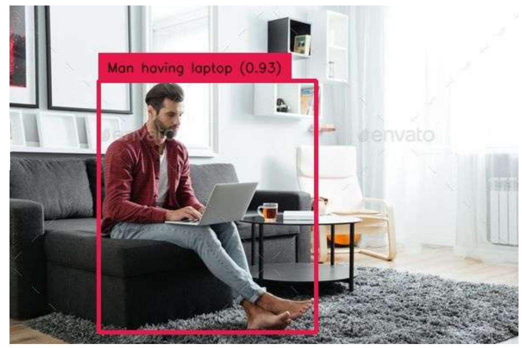
- **Device:** CPU (torch.device('cpu')).
- **Thresholds**: box_threshold≈0.2–0.25; NMS IoU≈0.6; WBF IoU≈0.35.
- **Data Flow (Output):** images in. /data/, results in timestamped ./results/ subfolders.
- **Graceful Degradation:** fallback to Soft-NMS if torchvision NMS unavailable;
- Skip re-ranking if **CLIP** import fails.

6. Experiments / Qualitative Results

- Demonstrated on scene queries like 'person working on a laptop', 'vendor selling vegetables'. (*in my system while running)*
- Re-ranking improves ambiguous cases; tiling helps with large images. Failure modes: ambiguous prompts, very small targets, extreme occlusions.

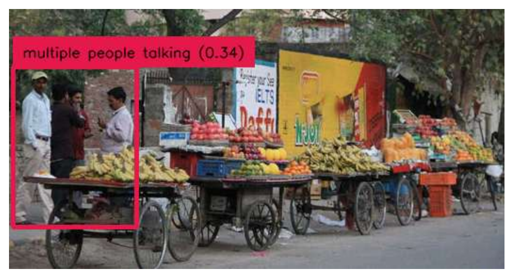
QUERY

Man having laptop



QUERY

Multiple people talking



7. Limitations and Future Work

- **CLIP** and **SAM** add CPU latency; quantization or smaller backbones could help.
- Negative prompts are heuristic; more principled joint constraints possible.
- Multi-object output and temporal consistency are not covered in this version.

8. Alternative Approaches Explored

1. **GroundingDINO only**
    - ***Why considered***: Simple baseline for fast text-to-box grounding.
    - ***Why not finalized***: Failed on complex queries and lacked precise masks.
2. **OWL-ViT**
    - ***Why considered***: Open-vocabulary detection with zero-shot potential.
    - ***Why not finalized***: Poor accuracy when compared to GroundingDino.
3. **Fine-tuning GroundingDINO**
    - ***Why considered***: Could improve accuracy on domain-specific queries.
    - ***Why not finalized***: Needed heavy compute and I don’t have GPU