LaMER: Language-Conditioned Multi-Expert Routing for Weakly Supervised Referring Expression Comprehension

Anonymous authors
Paper under double-blind review

Abstract

Referring Expression Comprehension (REC) localizes the image region described by a natural language expression, underpinning applications such as autonomous navigation, human-robot interaction, and medical image analysis. Under weak supervision, REC is challenged by the visual-linguistic semantic gap and the difficulty of inferring region-text correspondences from coarse image-level signals alone. Moreover, no single visual encoder captures all complementary properties—semantic alignment, fine-grained texture, and boundary awareness—while naively fusing multiple pretrained encoders yields heterogeneous, hard-to-reconcile representations. We propose LaMER, a weakly supervised REC framework that dynamically aggregates four frozen expert encoders through a language-conditioned Cross-Modal Routing Fusion module, adapting expert selection to each referring expression without instance-level bounding-box supervision. A Cell-Text Matching Head with Negative Anchor Augmentation, trained with a multi-objective loss combining contrastive, reconstruction, alignment, and load-balancing terms, further stabilizes optimization. LaMER achieves 77.67%, 62.81%, and 69.49% Acc@IoU$\geq$ 0.5 on RefCOCO, RefCOCO+, and RefCOCOg, surpassing DViN and WeakMCN by up to 16.49 and 14.49 points, with ablations validating each component.

1 Introduction

Understanding and localizing natural language references to visual content is a core requirement for intelligent systems operating in the real world. Referring Expression Comprehension (REC), the task of localizing the image region described by a natural language expression, underpins applications such as autonomous navigation (Deruyttere et al., 2019), human-robot interaction (Shridhar et al., 2020), and medical image analysis (Lei et al., 2023), where systems must identify the target object among multiple visually similar candidates.

REC remains fundamentally challenged by a visual-linguistic semantic gap between continuous visual representations and discrete language tokens, an absence of instance-level supervision, and the difficulty of integrating heterogeneous visual representations. These challenges intensify under weak supervision, where models learn solely from image-level signals, making the quality of visual representations the primary bottleneck: features must be both spatially discriminative and semantically aligned with the language query. Yet no single visual encoder simultaneously captures all complementary visual cues—semantic alignment, fine-grained texture, structure, and boundary awareness. Integrating multiple specialized pretrained encoders through a Mixture-of-Experts (MoE) paradigm (Shazeer et al., 2017) is a promising direction, but heterogeneous experts produce representations optimized for different objectives and feature distributions, making naive fusion ineffective; a learned language-conditioned router further risks collapsing onto a small subset of experts without appropriate load-balancing regularization.

We review existing REC methods in detail in Section 2. In brief, fully supervised approaches (Yu et al., 2018; Hu et al., 2017; Wang et al., 2019; Yang et al., 2020; Deng et al., 2021; Kamath et al., 2021; Choi et al., 2022) achieve strong performance but depend on costly instance-level annotations that limit scalability (Liu et al., 2019a), while recent weakly supervised methods (Jin et al., 2023;

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

Luo et al., 2024) instead pair open-vocabulary detectors (Gu et al., 2022; Cheng et al., 2024) with a single visual backbone. This reliance on one encoder limits the diversity of visual characteristics captured, leaving a gap in aggregating heterogeneous visual representations under weak supervision.

To address these limitations, we propose LaMER (Language-Conditioned Multi-Expert Routing for Weakly Supervised Referring Expression Comprehension), a weakly supervised REC framework that dynamically aggregates complementary visual representations from heterogeneous pretrained visual experts. Instead of a single visual backbone, LaMER leverages multiple frozen expert encoders and aggregates their outputs through a learned gating mechanism in a Cross-Modal Routing Fusion module and a Cell-Text Matching Head with Negative Anchor Augmentation, trained with a multi-objective optimization strategy that stabilizes expert routing and improves grounding performance without requiring bounding-box annotations.

Extensive experiments on RefCOCO, RefCOCO+, and RefCOCOg show that LaMER outperforms the state-of-the-art weakly supervised methods DViN and WeakMCN by up to 16.49 and 14.49 percentage points, with comprehensive ablations validating the effectiveness of each proposed component. The main contributions of this paper are summarized as follows:

- We propose LaMER, a weakly supervised referring expression comprehension framework that introduces a language-conditioned multi-expert routing mechanism for aggregating complementary visual representations from heterogeneous pretrained visual encoders.
- We design a *Cross-Modal Routing Fusion* module and an *Expert Feature Bank*, enabling learned weighted integration of heterogeneous visual representations.
- We introduce a *Cell-Text Matching Head* with Negative Anchor Augmentation and optimize the framework using a multi-objective training strategy that stabilizes expert routing and improves weakly supervised grounding without requiring bounding-box annotations.
- Extensive experiments on RefCOCO, RefCOCO+, and RefCOCOg demonstrate state-of-the-art performance among weakly supervised REC methods, while comprehensive ablation studies verify the effectiveness of each proposed component.

2 Related Work and Preliminaries

2.1 Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) extends detection beyond predefined categories to novel concepts described through natural language. Early semantic-transfer approaches were limited by weak visual-semantic alignment; large-scale vision-language pretraining (e.g., CLIP, ALIGN) improved this via contrastive learning on image-text pairs. Building on such models, ViLD (Gu et al., 2022) distills a frozen CLIP teacher into a detector; GLIP and RegionCLIP reformulate detection as phrase grounding; YOLO-World (Cheng et al., 2024) introduces a lightweight vision-language pathway for real-time inference. Despite these advances, OVD models remain designed for category-level recognition and are insufficient for compositional expressions involving attributes, spatial relationships, and context—so directly applying OVD representations to REC remains challenging.

2.2 Referring Expression Comprehension

Referring Expression Comprehension (REC) aims to localize the image region described by a natural language expression, requiring models to identify the intended instance among visually similar candidates. Existing approaches divide into fully supervised and weakly supervised paradigms.

Fully supervised REC. Early REC approaches relied on modular architectures that decomposed expressions into semantic components such as object categories, attributes, and spatial relationships (Yu et al., 2018; Hu et al., 2017), while graph-based methods modeled objects as nodes and exploited relational reasoning over scene structures (Wang et al., 2019; Yang et al., 2020). With Transformer-based vision-language architectures, approaches such as TransVG (Deng et al., 2021), MDETR (Kamath et al., 2021), and SeqTR (Choi et al., 2022) achieved strong performance through end-to-end cross-modal interaction and localization, but their dependence on expensive instance-level bounding-box annotations limits scalability to large-scale scenarios (Liu et al., 2019a).

Zero-shot and weakly supervised REC. To reduce dependence on annotations, studies investigate weakly supervised and zero-shot REC, learning grounding from image-level supervision or transferring it from pretrained vision-language models, formulating REC as region-text matching over open-vocabulary detector proposals (Gu et al., 2022; Cheng et al., 2024). RefCLIP (Jin et al., 2023), APL (Luo et al., 2024), DViN (Chen et al., 2025), and WeakMCN (Cheng et al., 2025) improve weakly supervised grounding through enhanced visual-language alignment and auxiliary learning; Chu et al. (2024) augment RefCLIP's anchor-matching backbone with cross-modality attention, while Mi et al. (2024) distill teacher-grounder knowledge under a ground-truth-proposal regime rather than LaMER's fully automatic candidate generation. QueryMatch (Chen et al., 2024) and AlignCAT (Wang et al., 2025) replace anchor-level matching with query-based representations from a Mask2Former detector; He et al. (2026) adopt a grounding DINO backbone with VLM-derived pseudo-labels; Liu et al. (2026) propose confidence-aware pseudo-label self-correction; and Xu et al. (2026) propose SVCR, routing multiple pretrained visual foundation models into structured relational features—the closest architectural precedent to LaMER's multi-expert routing.

However, most of these recent methods rely on a single visual backbone, limiting their ability to capture the diverse visual properties complex expressions require, and lack mechanisms to dynamically select complementary visual information per expression. LaMER addresses this gap via a language-conditioned routing mechanism over heterogeneous, frozen visual expert encoders, building on three established techniques—attention, mixture-of-experts routing, and contrastive learning—each mapped to a specific module or loss in Section 3; a self-contained review is provided in Appendix A.

3 Proposed Method

3.1 Overview

LaMER couples a YOLO11x-based visual ncoder (Khanam & Hussain, 2024) and a Bi-GRU language encoder with a Cross-Modal Routing Fusion module that assigns soft weights to four frozen expert encoders and dynamically mixes them with the backbone feature through a learnable residual gate, strengthening cross-modal alignment under weak supervision.

The fused representation is subsequently processed by the Cell-Text Matching Head, which performs region-text alignment through anchor-level contrastive matching. Negative Anchor Augmentation adds challenging negatives during training to improve discrimination, and LaMER is optimized with a multi-objective loss combining contrastive, reconstruction, alignment, load-balancing, and localization objectives, enabling stable routing without instance-level bounding-box annotations.

3.2 Research Challenges

Beyond the visual-linguistic semantic gap and the absence of instance-level supervision introduced in Section 1, realizing this design raises two further architecture-specific challenges:

1. **Visual Expert Heterogeneity:** The four frozen experts—SigLIP (image-text contrastive learning), DINOv2 (self-supervised learning), EfficientSAM (mask prediction), and ConvNeXt (supervised classification)—are pretrained under heterogeneous objectives and exhibit substantially different feature distributions, making simple fusion strategies such as averaging or concatenation ineffective.
2. **Dynamic Routing Stability:** Training a language-conditioned router jointly with the grounding objective may collapse it toward a single dominant expert; balancing expert utilization against the grounding objective requires carefully designed load-balancing losses, as overly strong regularization can interfere with the primary task.

3.3 Proposed System Architecture

As illustrated in Figure 1, the system consists of five components: the YOLO11x visual encoder, the Bi-GRU language encoder, the Expert Feature Bank, the Cross-Modal Routing Fusion module, and the Cell-Text Matching Head. The input image $I \in \mathbb{R}^{H\times W\times 3}$ and referring expression $Q = \{w_1, \ldots, w_T\}$ are fed in parallel into the visual and language encoders, extracting multi-scale visual features and a 512-dimensional context-aware language representation. These visual features are

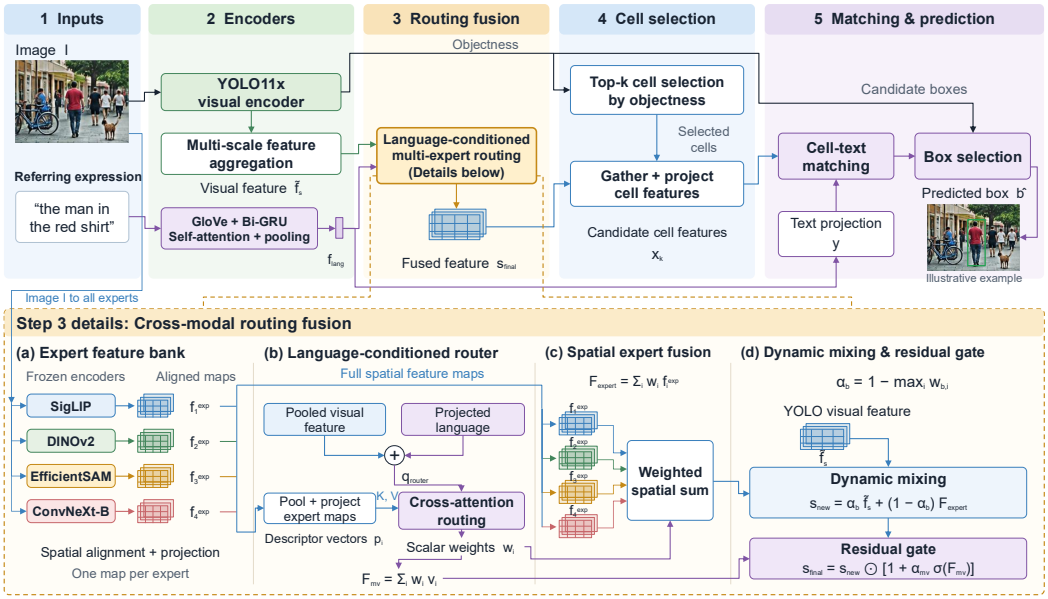


Figure 1: Overall architecture of the proposed LaMER framework.

enriched by four frozen experts in the Expert Feature Bank—SigLIP SO400M (Zhai et al., 2023), DINOv2 ViT-L/14 (Oquab et al., 2023), EfficientSAM (ViT-T) (Xiong et al., 2024), and ConvNeXt-B (Liu et al., 2022b)—each projected to a common 13×13, 1024-channel resolution. The Cross-Modal Routing Fusion module weights each expert's contribution by conditioning on a joint visual-linguistic query, producing a fused visual feature passed to the Cell-Text Matching Head, which selects the top-K most query-relevant anchor regions via objectness scores and produces the final bounding box $\hat{b} = (x, y, w, h)$ without bounding-box supervision during training.

3.3.1 Data Preprocessing

This stage converts raw image–text pairs from the RefCOCO, RefCOCO+, and RefCOCOg datasets (Yu et al., 2016; Mao et al., 2016; Kazemzadeh et al., 2014) into a format suitable for the encoding and grounding modules; consistent with the weakly supervised setting, ground-truth boxes are withheld during training and used only for evaluation. Images are resized to 416×416 pixels for the YOLO11x backbone, and each of the four expert encoders receives its own resolution- and normalization-matched view, since they were pretrained under different conventions; full per-branch preprocessing values are given in Appendix B. We tokenize each referring expression into GloVe (Pennington et al., 2014) word-index sequences, padded or truncated to $T_{\max} = 15$ tokens.

3.3.2 Feature Extraction

Three parallel branches process the preprocessed image and referring expression, producing the visual and linguistic representations that the Cross-Modal Routing Fusion module consumes.

YOLO11x (Khanam & Hussain, 2024) serves as the primary visual encoder, chosen for its ability to jointly produce region proposals and semantically rich multi-scale features within a single backbone. It builds hierarchical representations through C3k2 blocks with progressively larger receptive fields, an SPPF layer, and a C2PSA (Cross-Stage Partial with Spatial Attention) module at its final stage; the backbone is kept frozen except for C2PSA, which is selectively unfrozen so its spatial attention can adapt to the grounding task without disturbing the earlier general-purpose layers. The encoder outputs three feature maps at different spatial scales—$f_L \in \mathbb{R}^{B\times1024\times13\times13}$, $f_M \in \mathbb{R}^{B\times512\times26\times26}$, and $f_S \in \mathbb{R}^{B\times256\times52\times52}$—and objectness scores from the frozen detection heads serve as the anchor candidates used in Section 3.3.3. The referring expression is encoded with a Bidirectional GRU (Bi-GRU) over frozen GloVe (Pennington et al., 2014) word embeddings (300-dimension), chosen so that spatial modifiers common in referring expressions (e.g., "on

the left") are contextualized using information from both directions rather than only after the head noun. Each direction maintains a 256-dimension hidden state; the concatenated forward/backward outputs form a 512-dimension contextual token representation, further refined by three layers of multi-head self-attention and compressed by attention-based pooling into a global language feature $f_{\text{lang}} \in \mathbb{R}^{B\times 512}$. Following preprocessing (Section 3.3.1), each of the four frozen experts in the Expert Feature Bank — SigLIP SO400M (Zhai et al., 2023), DINOv2 ViT-L/14 (Oquab et al., 2023), EfficientSAM (ViT-T) (Xiong et al., 2024), and ConvNeXt-B (Liu et al., 2022b) — produces a complementary representation covering cross-modal semantic alignment, fine-grained self-supervised structure, boundary-aware localization, and local texture, respectively, each projected to 1024 channels and spatially aligned to 13×13 to match f_L.

3.3.3 Cross-Modal Routing Fusion

The Cross-Modal Routing Fusion module integrates the multi-scale YOLO11x features with the four expert representations under a joint visual-linguistic query, operating in four stages: multi-modal fusion, cross-attention routing, expert fusion with dynamic mixing, and top-k cell selection.

Multi-Modal Fusion. Before routing, a grounding projection adapts the YOLO11x feature maps to the task: f_L passes through a bottleneck projection ($1024 \to 512 \to 512 \to 1024$) to remap the C2PSA spatial attention features, while the medium- and fine-scale features pass through identity-initialized 1×1 convolutions for stable scale-specific adjustment. The three projected maps are integrated by a custom multi-scale fusion module, ***MultiScaleFusion***, following a Feature Pyramid Network design (Lin et al., 2017), combining an upsampling path that propagates high-level semantics to finer scales with a downsampling path that aggregates fine-grained detail into coarser scales:

$$\tilde{f}_s, \tilde{f}_m, \tilde{f}_l = \text{MultiScaleFusion}(f_L, f_M, f_S), \tag{1}$$

where $\tilde{f}_s \in \mathbb{R}^{B\times 1024\times 13\times 13}$ serves as the primary visual representation for routing. A joint visual-linguistic router query is then formed by adding the globally pooled $\tilde{f}_s$ to the projected language feature:

$$q_{\text{router}} = \text{AvgPool}(\tilde{f}_s) + W_{\text{lang}} f_{\text{lang}} \in \mathbb{R}^{B\times 1024}, \tag{2}$$

where $W_{\text{lang}} \in \mathbb{R}^{512\times 1024}$ projects the language feature into the routing space.

Cross-Attention Router. Since a static fusion strategy cannot reflect the differing usefulness of the 4 experts for grounding, a soft routing mechanism based on scaled dot-product attention learns continuous weights to all four experts simultaneously, keeping every expert active rather than committing to a hard selection. Each expert's feature map is globally pooled and projected into a shared descriptor $p_i \in \mathbb{R}^{B\times 1024}$, stacked into $P \in \mathbb{R}^{B\times 4\times 1024}$. With projections $q = W_Q q_{\text{router}}$, $K = W_K P$, $V = W_V P$, the routing weights and mixed-value representation follow scaled dot-product attention:

$$w = \text{softmax}\left(\frac{qK^\top}{\sqrt{d}}\right) \in \mathbb{R}^{B\times 4}, \quad F_{mv} = \sum_{i=1}^{4} w_i \cdot v_i \in \mathbb{R}^{B\times d}, \tag{3}$$

where w_i is the i-th expert's routing weight, and F_{mv} encodes expert selection and value blending.

Expert Fusion and Dynamic Mixing. The routing weights are applied to the full spatial feature maps of the four experts to produce a weighted expert representation $F_{\text{expert}} = \sum_{i=1}^{4} w_i \cdot f_i^{\text{exp}}$. This is blended with the YOLO11x backbone feature through a per-sample mixing coefficient derived from the peak routing weight:

$$\alpha_b = 1 - \max_i w_{b,i}, \qquad s_{\text{new}} = \alpha_b \cdot \tilde{f}_s + (1 - \alpha_b) \cdot F_{\text{expert}}, \tag{4}$$

so that a concentrated routing distribution shifts the blend toward the dominant expert, while a flatter distribution retains more of the YOLO11x backbone feature. A learnable residual gate conditioned on F_{mv} then applies a final channel-wise adjustment:

$$s_{\text{final}} = s_{\text{new}} \cdot \left(1 + \alpha_{mv} \cdot \sigma(F_{mv})\right), \tag{5}$$

where α_{mv} is a scalar initialized to zero, so the gate begins as an identity mapping and only gradually contributes as training progresses, and σ is the sigmoid function.

Top-k Cell Selection. After fusion, the $13 \times 13 = 169$ grid cells are ranked by the mean objectness of their four anchor proposals from the YOLO11x detection head ($13 \times 13 \times 4 = 676$ proposals in total), and the top $k = 17$ cells are selected:

$$A_{\text{sel}} = \text{TopK}\left(\{\bar{o}_c\}_{c=1}^{169}\right), \tag{6}$$

where $\bar{o}_c$ is the mean objectness of the anchors in cell c and A_{sel} is the set of selected cells. The spatial features at the corresponding grid positions in s_{final} are extracted and projected to 512 dimensions, yielding $X \in \mathbb{R}^{B\times 17\times 512}$, while the language feature f_{lang} is projected to $Y \in \mathbb{R}^{B\times 1\times 512}$. The pair (X, Y) is then passed to the Cell-Text Matching Head.

3.3.4 Cell-Text Matching Head

The Cell-Text Matching Head receives the top-k anchor feature matrix $X \in \mathbb{R}^{B\times 17\times 512}$ and the global language feature $Y \in \mathbb{R}^{B\times 1\times 512}$ from Top-k Cell Selection, and produces the final grounding prediction through contrastive cell-text matching.

A pairwise similarity score is first computed between each of the 17 anchor representations and the language query via scaled dot-product:

$$S = \frac{XY^\top}{\sqrt{d}} \in \mathbb{R}^{B\times 17}, \tag{7}$$

where $d = 512$. To increase discriminative difficulty, Negative Anchor Augmentation appends the top-2 highest-scoring anchors from other samples within the same batch as hard negatives, forcing the model to distinguish the target from semantically plausible candidates drawn from entirely different images rather than trivially low-scoring distractors.

At inference, the prediction is the anchor with the highest similarity score:

$$\hat{k} = \arg\max_{k\in\{1,\ldots,17\}} S_k, \tag{8}$$

whose coordinates (x, y, w, h) are decoded from the YOLO11x detection head as the final predicted bounding box $\hat{b}$.

3.3.5 Loss Function

The primary weakly supervised signal is the anchor-prompt contrastive loss $\mathcal{L}_{\text{anchor}}$, aligning the top-17 anchor visual features with the language query via the augmented InfoNCE-style (van den Oord et al., 2018) objective above.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ adds cross-modal regularization: a three-layer MLP decoder reconstructs the mean of the top-17 anchor visual features back into the language feature space:

$$\mathcal{L}_{\text{recon}} = \frac{1}{17}\sum_{k=1}^{17} \left\|\text{Decoder}(x_k) - f_{\text{lang}}\right\|^2, \tag{9}$$

encouraging selected anchors to retain semantic consistency with the query even after expert fusion.

To discourage expert collapse, a load-balancing loss $\mathcal{L}_{\text{sparse}}$ penalizes deviation of the batch-averaged routing distribution from a uniform prior:

$$\mathcal{L}_{\text{sparse}} = \text{MSE}\left(\bar{w}, \tfrac{1}{4}\mathbf{1}\right), \qquad \bar{w} = \frac{1}{B}\sum_{b=1}^{B} w_b, \tag{10}$$

where $w_b \in \mathbb{R}^4$ is the b-th sample's routing weight vector.

The text-visual alignment loss $\mathcal{L}_{\text{align}}$ gives the routing module a more direct cross-modal signal, applying InfoNCE (van den Oord et al., 2018) at $\tau = 0.5$ between projections of the fused expert features and the language feature:

$$\mathcal{L}_{\text{align}} = \text{InfoNCE}\left(\text{Proj}_{\text{expert}}(F_{\text{expert}}), \text{Proj}_{\text{text}}(f_{\text{lang}}); \tau = 0.5\right), \tag{11}$$

where both projectors are two-layer MLPs mapping to $\mathbb{R}^{128}$; without this loss the router would receive cross-modal gradient only indirectly through $\mathcal{L}_{\text{anchor}}$.

Table 1: Statistics of the benchmark datasets.

Dataset	Images	Objects	Expressions	Train / Val / Test
RefCOCO	19,994	50,000	142,210	120.6k / 10.8k / 10.7k
RefCOCO+	19,992	49,856	141,564	120.1k / 10.7k / 10.6k
RefCOCOg	25,799	49,822	95,010	80.5k / 4.8k / 9.6k

Finally, the box regression loss $\mathcal{L}_{\text{box}}$ provides optional spatial supervision: a soft-attention weighted average of the selected anchors initializes a prediction refined by a two-layer MLP residual head, supervised with Generalized IoU (Rezatofighi et al., 2019):

$$\mathcal{L}_{\text{box}} = \mathcal{L}_{\text{GIoU}}(\hat{b}_{\text{pred}}, b_{\text{gt}}). \tag{12}$$

The overall training objective is a weighted sum of the five terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{recon}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}}. \tag{13}$$

$\mathcal{L}_{\text{box}}$ and its refinement head support optional supervised fine-tuning, but we set $\lambda_{\text{box}} = 0$ in all reported experiments, so ground-truth boxes never contribute gradient signal, and predictions are decoded directly from the selected anchor, consistent with the weakly supervised setting.

4 Experiments

4.1 Datasets

We evaluate on three standard REC benchmarks: RefCOCO and RefCOCO+ (Yu et al., 2016), both built on MS-COCO train2014 images via the interactive ReferItGame protocol (Kazemzadeh et al., 2014) and split following the UNC convention (testA: person queries, testB: non-person queries), and RefCOCOg (Mao et al., 2016), collected non-interactively via Amazon Mechanical Turk (Amazon Web Services, 2005) with longer, more descriptive expressions (8.4 words on average vs. 3.5 for RefCOCO), split following the UMD convention. RefCOCO+ additionally prohibits spatial terms (e.g., "left"/"right"), isolating appearance-based grounding. Table 1 summarizes dataset statistics; Appendix B provides full per-dataset details.

4.2 Evaluation Metrics

Following standard practice in REC evaluation, we adopt Accuracy at IoU threshold 0.5 (Acc@IoU≥ 0.5) as the evaluation metric: a prediction is correct if the Intersection over Union between the predicted box $\hat{b}$ and ground-truth box b^* exceeds 0.5,

$$\text{IoU}(\hat{b}, b^*) = \frac{\text{Area}(\hat{b} \cap b^*)}{\text{Area}(\hat{b} \cup b^*)}, \tag{14}$$

and the overall metric is the fraction of the N test samples satisfying this threshold.

4.3 Experimental Setup

LaMER is optimized with Adam (Kingma & Ba, 2015) under a warmup-cosine schedule for 55 epochs (batch size 32, gradient accumulation over 2 steps), with the YOLO11x backbone and GloVe embeddings frozen except for the C2PSA attention block, which is selectively fine-tuned. Full hyperparameters and hardware configuration are provided in Appendix B.

4.4 Experimental Results and Analysis

4.4.1 Comparison with State-of-the-Art Methods

Table 2 reports Acc@IoU≥ 0.5 for LaMER against representative weakly supervised REC methods, grouped by proposal source: *GT Proposals* methods rely on ground-truth region candidates at inference, while *Det Proposals* methods, including LaMER, generate candidates automatically via

Table 2: Comparison with state-of-the-art methods on REC datasets.

Method	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val
GT Proposals							
VC (Niu et al., 2018)	–	33.3	30.1	–	34.6	–	30.3
ARN (Liu et al., 2019a)	38.1	36.4	36.5	34.5	36.4	–	39.6
KPRN (Liu et al., 2019b)	36.3	35.3	37.7	37.2	36.1	–	38.4
DTWREG (Sun et al., 2021)	39.2	41.1	37.7	39.2	40.0	–	43.2
EARN (Liu et al., 2022a)	38.1	38.3	38.6	37.5	37.6	–	45.3
Adaptive KD (Mi et al., 2024)†	39.8	39.9	40.2	40.1	40.1	41.1	48.0
Det Proposals							
VC (Niu et al., 2018)	–	32.7	27.2	–	34.7	–	29.7
ARN (Liu et al., 2019a)	32.2	35.3	30.3	32.8	34.4	–	33.1
IGN (Zhang et al., 2020)	34.8	37.6	32.6	34.3	36.9	–	34.9
DTWREG (Sun et al., 2021)	38.4	39.5	37.0	38.9	39.9	–	42.5
RefCLIP (Jin et al., 2023)	60.4	58.6	57.1	40.4	40.5	–	47.9
QueryMatch (Chen et al., 2024)	66.0	66.0	65.5	44.8	46.7	41.5	48.5
APL (Luo et al., 2024)	64.5	61.9	63.6	42.7	42.8	–	50.2
CMA (Chu et al., 2024)	63.9	61.4	63.8	41.8	42.1	40.4	48.4
DViN (Chen et al., 2025)	67.7	70.9	59.4	52.5	57.5	–	55.0
WeakMCN (Cheng et al., 2025)	69.2	70.8	63.4	51.9	57.4	43.3	55.0
AlignCAT (Wang et al., 2025)	69.0	70.3	66.6	47.2	52.2	41.9	54.7
NRG (He et al., 2026)	69.6	73.2	67.7	56.0	64.5	46.2	58.3
SVCR (Xu et al., 2026)	71.1	72.6	65.6	54.2	60.2	44.6	57.0
SVCR (Xu et al., 2026) (YOLOv5)	74.5	76.9	70.5	56.4	64.2	48.7	60.4
LaMER (Ours)	**77.7**	**79.1**	**75.9**	**62.8**	**67.9**	**57.0**	**69.5**

† Adaptive KD trains on ground-truth-derived triads, an easier setting than the Det Proposals methods below; listed under GT Proposals for like-for-like grouping, not as a weaker model. LaMER's RefCOCOg result is on the UMD validation split; baseline RefCOCOg results are quoted from the original papers, and WeakMCN results from its official repository. LaMER's RefCOCOg result is on the UMD validation split; other RefCOCOg results are as reported in the original papers, and WeakMCN results are from its official repository.

a pretrained detector without any annotation—a harder setting, since the candidate set is not guaranteed to contain the correct region. Despite this handicap, LaMER leads all Det Proposals baselines and surpasses every GT Proposals method across all three benchmarks.

Against the strongest prior Det Proposals method, WeakMCN, LaMER's gains widen as the task grows harder: +8.45/+8.33/+12.45 on RefCOCO val/testA/testB; +10.88/+10.47/+13.73 on RefCOCO+, which prohibits spatial terms and forces appearance-based reasoning, indicating the Expert Feature Bank's texture- and structure-sensitive experts supply discriminative information single-encoder methods lack; and the largest margin overall, +14.49 over WeakMCN and +19.27 over APL, on RefCOCOg, whose longer expressions combine appearance and relational cues—suggesting the Cross-Attention Router adapts its expert weighting more effectively to multi-clause queries.

Query-based matching (QueryMatch (Chen et al., 2024), AlignCAT (Wang et al., 2025)) and enhanced anchor-matching (CMA (Chu et al., 2024)) all trail LaMER by 8 to 26 points, indicating that stronger cross-modal alignment alone does not close the gap left by a single visual backbone. NRG (He et al., 2026), fine-tuning Grounding DINO via LoRA with VLM-derived pseudo-labels, trails by 3.4 to 11.2 points. The closest competitor is SVCR (Xu et al., 2026), routing multiple foundation models into relational representations: under its default YOLOv3 backbone, SVCR trails by 6.5 to 12.5 points, narrowing to +2.16 on RefCOCO testA with a stronger YOLOv5 backbone. Even so, LaMER outperforms both SVCR and NRG on every split under every backbone setting.

4.5 Ablation Studies

We ablate each component on RefCOCO and RefCOCO+ under Acc@IoU≥ 0.5. The expert-combination and encoder-count studies are reported in Appendix C; here we focus on the cumulative component ablation.

4.5.1 Cumulative Component Ablation

Table 3 adds each architectural component incrementally. The baseline of YOLO11x with four statically fused experts reaches 74.11% and 56.22% on the RefCOCO and RefCOCO+ val splits. The Cross-Attention Router yields the largest single gain (+2.51 and +4.53 points), confirming that learned expert weighting is central to exploiting complementary experts. Multi-Scale Fusion further improves RefCOCO val by +0.83, though RefCOCO+ val regresses slightly (-0.57) at this stage; the Bottleneck grounding projection recovers this, adding +2.09 on RefCOCO+ val, so the

Table 3: Cumulative ablation study of LaMER.

Method	RefCOCO			RefCOCO+		
	val	testA	testB	val	testA	testB
YOLO11x + 4 Experts	74.1	74.2	72.9	56.2	61.3	50.9
+ Cross-Attn Router	76.6	77.9	73.6	60.8	65.9	55.4
+ MultiScale Fusion	77.5	78.3	74.2	60.2	65.5	53.8
+ Bottleneck Proj.	77.6	78.7	75.4	62.3	67.6	56.2
+ Unfreeze C2PSA	**77.7**	**79.1**	**75.9**	**62.8**	**67.9**	**57.0**



(a) woman in black sweatshirt standing with brown hair

(b) woman in dark blue jacket with long blond hair

(c) far left

(d) girl glasses black shirt front



(e) second row man yellow tie on the left

(f) center man white hair yellow tie

(g) bottom right girl half out

(h) man right striped tie

Figure 2: Same image, different queries. Blue: LaMER prediction; green dashed: annotated target. Panels (a)–(d) are from a ReferItGame test image and (e)–(h) from a RefCOCO testA image. The queries in (a)–(c) and (e)–(h) lead LaMER to seven different people; in (d) the target is among the candidate boxes, but LaMER returns the box of (a).

two components are complementary. Selectively unfreezing C2PSA provides a final, smaller gain (+0.40/+0.30 testA), consistent with its role of limited task-specific adaptation.

4.6 Qualitative Analysis

Figure 2 pairs two images with eight referring expressions. The predictions follow the queries: across the two scenes, the model resolves seven different people, including distinctions that rest on clothing color, hair color, relative position, and partial visibility. Panel (d) is the one failure, where the annotated target is among the candidate boxes but the model returns the box of (a).

5 Conclusion and Future Work

This work presents LaMER, a weakly supervised REC framework that aggregates complementary visual representations in a query-adaptive manner without instance-level supervision. It couples a YOLO11x visual encoder and a Bi-GRU language encoder with an Expert Feature Bank of four frozen encoders via a Cross-Modal Routing Fusion module and a Cell-Text Matching Head with Negative Anchor Augmentation. Trained end-to-end with a multi-objective loss, LaMER achieves 77.67%, 62.81%, and 69.49% Acc@IoU$\geq$ 0.5 on RefCOCO, RefCOCO+, and RefCOCOg, surpassing DViN and WeakMCN by up to +16.49% and +14.49%, with ablations confirming each component's contribution. Future work includes extending LaMER to weakly supervised Referring Expression Segmentation and exploring adaptive sparse expert routing to reduce inference overhead.

REFERENCES

Amazon Web Services. Amazon mechanical turk. https://www.mturk.com, 2005.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Shengxin Chen, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, and Rongrong Ji. Querymatch: A query-based contrastive learning framework for weakly supervised visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pp. 4177–4186, 2024.

Xiaofu Chen, Yaxin Luo, Gen Luo, Jiayi Ji, Henghui Ding, and Yiyi Zhou. Dvin: Dynamic visual routing network for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14347–14357, 2025.

Silin Cheng, Yang Liu, Xinwei He, Sebastien Ourselin, Lei Tan, and Gen Luo. Weakmcn: Multi-task collaborative network for weakly supervised referring expression comprehension and segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 9175–9185, 2025.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Seokeon Choi, Sunghun Park, Sangdoo Kim, and Nojun Kwak. Seqtr: Sequence transformer for end-to-end text-to-image grounding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Ting-Yu Chu, Yong-Xiang Lin, Ching-Chun Huang, and Kai-Lung Hua. Enhancing anchor-based weakly supervised referring expression comprehension with cross-modality attention. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 2767–2782, 2024.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2088–2098, 2019. doi: 10.18653/v1/D19-1215.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Yunlong He, Ziqi Gu, Tong Zhang, Zhen Cui, and Chunyan Xu. From noisy candidates to reliable grounding in weakly-supervised referring expression comprehension. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, pp. 317–326, 2026. doi: 10.1145/3805622.3810686.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4418–4427, 2017.

Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2681–2690, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, 2021.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolge, Laura Xiao, Tete Wan, Laura Gustafson, Sean Hälfär, Nace Wu, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Zhihao Lei, Yankai Wei, Sheng Zhang, Suraj Jha, Bhavya Bhatt, and Petr Motlicek. Medical phrase grounding with region-phrase context contrastive alignment. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023. doi: 10.1007/978-3-031-43904-9_1.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2611–2620, 2019a.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, pp. 539–547, 2019b.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3003–3018, 2022a.

Yang Liu, Jiahua Zhang, Yao Wu, Zheng Zhao, Qingchao Chen, and Yuxin Peng. Confidence-aware pseudo-label self-correction for weakly supervised visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):8901–8918, 2026.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022b.

Yaxin Luo, Jiayi Ji, Xiaofu Chen, Yuxin Zhang, Tianhe Ren, and Gen Luo. Apl: Anchor-free prompt learning for one-stage weakly supervised referring expression comprehension. In *European Conference on Computer Vision (ECCV)*, pp. 198–215, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jinpeng Mi, Stefan Wermter, and Jianwei Zhang. Adaptive knowledge distillation and integration for weakly supervised referring expression comprehension. *Knowledge-Based Systems*, 286:111437, 2024.

Yulei Niu, Hanwang Zhang, Zhiwu Lu, and Shih-Fu Chang. Variational context: Exploiting visual and textual context for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1992–2001, 2018.

Nabil Omi, Siddhartha Sen, and Ali Farhadi. Load balancing mixture of experts with similarity preserving routers. arXiv preprint arXiv:2506.14038, 2025. URL `https://arxiv.org/abs/2506.14038`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)*, 2017.

Mohit Shridhar, Dixant Mittal, and David Hsu. INGRESS: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2–3):217–232, 2020. doi: 10.1177/0278364919897133.

Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y. Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4189–4195, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1960–1968, 2019.

Yidan Wang, Chenyi Zhuang, Wutao Liu, Pan Gao, and Nicu Sebe. Aligncat: Visual-linguistic alignment of category and attribute for weakly supervised visual grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Yunyang Xiong, Balaji Varadarajan, Lemeng Wu, Xiaoyu Zhu, Fanyi Xiao, Chenchen Zhu, Zechun Dai, Dilin Wang, Zhichao Liang, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13111–13121, 2024.

Lian Xu, Mohammed Bennamoun, Farid Boussaid, Hamid Laga, Yulan Guo, and Dan Xu. Learning structured visual compositional representations for weakly supervised referring expression comprehension.

arXiv:2607.04638, 2026.

Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9952–9962, 2020.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, 2016.

Licheng Yu, Hao Shi, Frank Yu, Xin Tian, Zhelin Huang, Varun K. Nagaraja, Yejin Choi, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1307–1315, 2018.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18123–18134, 2020.

A Background

This appendix reviews the three established techniques LaMER builds on – attention mechanisms, mixture-of-experts routing, and contrastive learning – and how each maps onto a specific module or loss described in Section 3.

Attention Mechanisms. Attention mechanisms model dependencies within or between sequences by computing relevance scores between queries and key-value pairs, dispensing with recurrence to enable parallel computation (Vaswani et al., 2017). Scaled Dot-Product Attention computes a weighted sum of values based on query-key similarity, while Multi-Head Attention runs several such operations in parallel across distinct learned subspaces, allowing the model to jointly attend to different types of relationships within the input. LaMER uses both mechanisms directly: scaled dot-product attention forms the basis of the Cross-Attention Router (Section 3.3.3), and multi-head self-attention refines the language representation (Section 3.3.1). Several of LaMER's frozen expert encoders also build on the closely related Vision Transformer architecture (Dosovitskiy et al., 2021): the Segment Anything Model (Kirillov et al., 2023) uses a ViT image encoder pretrained via Masked Autoencoders (He et al., 2022) to produce spatially rich embeddings, while DINO-ViT (Caron et al., 2021) uses self-supervised distillation to learn structural, label-free visual representations. LaMER leverages both properties through its EfficientSAM and DINOv2 experts, respectively.

Mixture of Experts. To enhance model capacity and adaptability without a proportional increase in computational cost, the Mixture of Experts (MoE) paradigm (Shazeer et al., 2017) is widely utilized. An MoE layer consists of a set of N expert networks $\{E_i\}_{i=1}^{N}$ and a gating network G, which dynamically routes inputs to different experts for conditional computation. For an input x, the output is a weighted sum of the experts' outputs, typically using a Top-K routing strategy where only the k experts with the highest gating scores are activated:

$$y = \sum_{i=1}^{N} G(x)_i E_i(x) \tag{15}$$

$$G(x) = \text{Softmax}(\text{TopK}(H(x), k)) \tag{16}$$

where $G(x)_i$ indicates the gating weight for the i-th expert, $E_i(x)$ is the output of the i-th expert, and $H(x)$ represents the raw logits produced by the gating network.

Load Balancing. During MoE training, a frequently encountered issue is expert collapse, where the gating mechanism disproportionately assigns most inputs to a small subset of experts, resulting in the remaining experts being rarely activated. To alleviate this imbalance and promote more even expert participation, an auxiliary load balancing loss is introduced and defined as follows:

$$L_{aux} = N \sum_{i=1}^{N} f_i \cdot P_i \tag{17}$$

where N represents the total number of experts. In this formulation, f_i indicates the empirical fraction of samples routed to expert i, while P_i denotes the average routing probability assigned to expert i by the gating network:

$$f_i = \frac{1}{B} \sum_{j=1}^{B} \mathbb{1}[i \in T_j] \tag{18}$$

$$P_i = \frac{1}{B} \sum_{j=1}^{B} G(x_j)_i \tag{19}$$

where T_j refers to the set of top-k experts chosen for input x_j, $G(x_j)_i$ is the gating network's predicted probability of selecting expert i given x_j, and B is the batch size. By minimizing L_{aux}, the model is encouraged to distribute inputs more evenly across experts, thereby improving overall expert utilization and reducing the risk of collapse (Omi et al., 2025). While Eq. (17) presents the canonical load-balancing formulation, Section 3.3.3 introduces an MSE-based variant (Eq. (10)) tailored to LaMER's four-expert, per-sample routing setting.

Contrastive Learning. Contrastive learning is a representation learning paradigm that structures an embedding space by pulling semantically related sample pairs closer together while pushing apart unrelated pairs. In multimodal learning settings, this approach is commonly used to align latent representations across different modalities. By constructing paired samples from different modalities or views, the model learns to map heterogeneous inputs into a shared embedding space, enabling semantic correspondence between representations.

A widely adopted objective function for contrastive learning is the InfoNCE loss. Given a batch containing N paired samples, the loss for the i-th pair is defined as:

$$L_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i^{(1)}, z_i^{(2)})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i^{(1)}, z_j^{(2)})/\tau)} \tag{20}$$

where $z_i^{(1)}$ and $z_i^{(2)}$ denote the embeddings of the two corresponding views or modalities in the i-th paired sample, $\text{sim}(\cdot)$ typically represents cosine similarity, and τ is a temperature hyperparameter that controls the smoothness of the similarity distribution. This objective increases the similarity between matched pairs while relatively decreasing the similarity to mismatched pairs within the batch, thereby encouraging the learning of discriminative and well-aligned representations.

B Implementation Details

B.1 Dataset Details

RefCOCO (Yu et al., 2016) is built on the MS-COCO train2014 image collection, collected through the ReferItGame (Kazemzadeh et al., 2014) platform, an interactive two-player game in which one annotator describes a target object while the other identifies it. The dataset contains 142,210 referring expressions over 50,000 annotated objects across 19,994 images, with an average expression length of 3.5 words. Following the standard UNC split, the data is divided into 120,624 training, 10,834 validation, 5,657 testA, and 5,095 testB expressions. TestA contains expressions referring exclusively to people, while testB covers all other object categories.

RefCOCO+ (Yu et al., 2016) shares the same images and collection protocol as RefCOCO but explicitly prohibits spatial location terms such as "left" or "right," requiring annotators to rely purely

on appearance-based cues such as color, texture, and category. The dataset contains 141,564 expressions over 49,856 objects across 19,992 images, with an average query length of 3.53 words, and follows the same UNC split structure as RefCOCO: 120,191 training, 10,758 validation, 5,726 testA, and 4,889 testB expressions.

RefCOCOg (Mao et al., 2016) is also built on MS-COCO train2014 images, but is collected non-interactively through Amazon Mechanical Turk (Amazon Web Services, 2005) rather than a two-player game, resulting in longer and more descriptive expressions. The dataset contains 95,010 expressions over 49,822 objects across 25,799 images , with an average expression length of 8.4 words. Unlike the other two datasets, RefCOCOg does not use a testA/testB split; following the UMD partition, it uses a single, undivided test set that ensures no image appears in more than one subset.

B.2 Experimental Environment

All experiments are conducted on a server equipped with two NVIDIA GeForce RTX 5090 GPUs (32 GB GDDR7X each), an AMD EPYC 9654 96-core CPU, and 256 GB of system RAM, running Ubuntu 22.04 LTS with CUDA 13.0 and PyTorch 2.11.0.

Table 4: Experimental environment configuration.

Component	Specification
Operating System	Ubuntu 22.04 LTS
CPU	AMD EPYC 9654 (96 cores)
System Memory	256 GB
GPU	2× RTX 5090 (32 GB GDDR7X)
CUDA Version	13.0
DL Framework	PyTorch 2.11.0

B.3 Training Hyperparameters

The model is optimized using Adam (Kingma & Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$) with a warmup cosine annealing learning rate schedule: a 5-epoch warmup from 10^{-7}, a peak learning rate of 2×10^{-4}, and a minimum of 10^{-6}. Training runs for 55 epochs, determined empirically via validation performance on RefCOCO, with a total batch size of 32 (16 per GPU) and gradient accumulation over 2 steps (effective batch size 64). The maximum query length is 15 tokens for RefCOCO and RefCOCO+, and 20 for RefCOCOg. Gradient clipping (max norm 0.15) and an Exponential Moving Average (EMA) of model weights are applied to stabilize and improve generalization, and automatic mixed precision (AMP) is used throughout. The YOLO11x backbone and GloVe embeddings are initialized from pretrained weights and kept frozen, with the exception of the C2PSA attention block at layer 10 of the YOLO11x backbone, which is partially fine-tuned to adapt to the grounding task.

B.4 Per-Branch Preprocessing

Input images are resized to 416×416 pixels and normalized with ImageNet statistics (mean $= [0.485, 0.456, 0.406]$, std $= [0.229, 0.224, 0.225]$) to match the distribution expected by the pretrained YOLO11x backbone. Because the four expert encoders in the Expert Feature Bank were pretrained under different resolutions and normalization conventions, each requires its own preprocessing branch: SigLIP SO400M receives 384×384 input, re-normalized to its own statistics (mean $=$ std $= [0.5, 0.5, 0.5]$); with a 14×14 patch size, this yields $27 \times 27 = 729$ patch tokens, later pooled to the target 13×13 grid. DINOv2 ViT-L/14 receives 364×364 input under standard ImageNet normalization. EfficientSAM (ViT-T) receives 1024×1024 input, also under ImageNet normalization. ConvNeXt-B operates directly on the same 416×416 ImageNet-normalized image as YOLO, requiring no additional resizing, since its cumulative $32\times$ downsampling stride natively produces 13×13 feature maps. Each referring expression is tokenized and mapped to GloVe (Pennington et al., 2014) word-index sequences, padded or truncated to a maximum length of $T_{\max} = 15$ tokens (20 for RefCOCOg).

C Additional Ablation Studies

This appendix reports two additional ablation studies referenced in Section 4, both conducted on RefCOCO and RefCOCO+ under Acc@IoU≥ 0.5.

C.1 Effect of Expert Feature Combinations

Table 5: Comparison of different mixtures of visual features on RefCOCO and RefCOCO+.

	RefCOCO			RefCOCO+		
Expert Encoder	val	testA	testB	val	testA	testB
YOLO11x (Base)	67.2	65.4	70.9	38.1	37.7	39.4
CLIP	37.5	44.1	30.2	36.6	41.8	27.0
SigLIP-ViT	49.3	55.7	46.2	39.7	46.6	31.4
DINOv2 ViT-B	59.1	59.5	56.3	40.5	48.0	34.1
DINOv2 ViT-L	60.3	61.6	59.6	42.5	49.0	35.8
ConvNeXt-B	29.4	34.7	24.0	29.2	34.2	24.0
EffSAM-ViT	45.9	48.6	45.4	31.5	34.7	29.5
+ DINOv2	74.2	75.8	75.0	56.5	59.1	52.0
+ DINO + Sig	76.4	77.3	75.6	58.4	62.9	53.3
+ D + Sig + SAM	76.9	78.4	75.8	59.5	62.8	54.1
LaMER (Ours)	**77.7**	**79.1**	**75.9**	**62.8**	**67.9**	**57.0**

Table 5 evaluates individual and combined expert configurations. Used in isolation, every expert encoderunderperforms the YOLO11x baseline on RefCOCO (67.21% val), with DINOv2 ViT-L achieving the strongest single-encoder result at 60.30%; CLIP alone yields only 37.51%, indicating that a single cross-modal encoder is insufficient for the fine-grained localization required under weak supervision. Progressively combining experts with YOLO11x yields consistent, substantial gains: adding DINOv2 ViT-L alone raises RefCOCO val from 67.21% to 74.23%, and subsequent additions of SigLIP and EfficientSAM further improve performance to 76.37% and 76.87%. Including ConvNeXt-B as the fourth expert reaches the full LaMER configuration at 77.67% on RefCOCO val and 62.81% on RefCOCO+ val, indicating that each expert contributes incremental, complementary information rather than redundant signal.

C.2 Ablation of Visual Encoder Count

Table 6: Ablation of the number of visual experts.

	RefCOCO			RefCOCO+		
Top-N	val	testA	testB	val	testA	testB
Top-1	38.1	38.6	38.3	29.2	29.2	26.6
Top-2	75.5	76.4	74.0	59.3	64.9	53.2
Top-3	77.1	78.8	75.5	61.0	67.5	55.1
Top-4	**77.7**	**79.1**	**75.9**	**62.8**	**67.9**	**57.0**

Table 6 varies the number of experts retained by the router at inference. Restricting the model to a single expert collapses performance to 38.06% and 29.15% on RefCOCO and RefCOCO+ val, far below any multi-expert configuration, confirming that cross-expert complementarity, rather than any individual encoder, drives the routing mechanism's effectiveness. Expanding to Top-2 recovers most of the gap (75.48% and 59.27%), and performance continues to improve monotonically through Top-3 and Top-4, with larger gains on RefCOCO+ than RefCOCO from Top-2 onward . This monotonic trend indicates that soft attention-based routing continues to extract complementary information as more experts are made available, without the diminishing returns typically associated with hard, single-expert selection.

C.3 Routing Weight Analysis

This appendix reports the routing weights behind the qualitative results of Section 4.6 and measures how much they contribute to the predictions.

Table 7: IoU with the annotated target and router weights over the four experts, for panels (a)–(h) of Figure 2.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
IoU	0.72	0.82	0.90	0.00	0.97	0.87	0.83	0.90
SigLIP	0.29	0.31	0.27	0.30	0.28	0.26	0.27	0.26
DINOv2	0.17	0.10	0.29	0.15	0.18	0.14	0.25	0.14
EfficientSAM	0.49	0.53	0.31	0.50	0.50	0.58	0.44	0.57
ConvNeXt	0.05	0.06	0.12	0.05	0.04	0.02	0.04	0.03

Table 7 gives the weights for the eight panels of Figure 2. They stay within narrow bands—SigLIP $[0.26, 0.31]$, DINOv2 $[0.10, 0.29]$, EfficientSAM $[0.31, 0.58]$, ConvNeXt $[0.02, 0.12]$—averaging $[0.28, 0.18, 0.49, 0.05]$, and EfficientSAM is the highest-weighted expert in all eight. Two quantities inherit this stability: the mixing coefficient $\alpha_b = 1 - \max_i w_i$ of Eq. (4) stays in $[0.42, 0.69]$, and the mean utilization is far from the uniform prior of Eq. (10), so $\mathcal{L}_{\text{sparse}}$ moderates the imbalance without removing it, while still keeping every expert active.