Comprehension of Multilingual Expressions Referring to Target Objects in Visual Inputs

Francisco Reis Nogueira[1] · Bruno Martins[1] · Alexandre Bernardino[2]

[1] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal [2] ISR-Lisboa, LARSyS, Instituto Superior Técnico, Universidade de Lisboa, Portugal

{francisco.r.nogueira, bruno.g.martins, alexandre.bernardino}@tecnico.ulisboa.pt

1 Motivation & Contributions

Referring Expression Comprehension (REC) localizes an object in an image from a natural-language description. Despite rapid progress in this problem, benchmarks and models remain almost exclusively **English-centric**.

- **Unified multilingual dataset:** 12 English REC benchmarks were systematically expanded to **10 languages** via machine translation, automatic quality estimation, and image-aware re-translation (≈8M expressions).
- **Attention-anchored architecture:** a compact model based on a natively multilingual **SigLIP2** encoder that derives a coarse bounding box *from its own cross-attention map*, and refines it with learned residuals.
- **Competitive & consistent:** Results point to 91.3% IoU@50 on the RefCOCO (English) dataset, with only 2-4 points of degradation on Romance languages, using a 469M-parameter model with ≈35M trainable parameters.

2 Unified Multilingual Dataset

12	10	≈8M	177.6K	336.9K
source benchmarks	languages	referring expressions	images	annotated objects

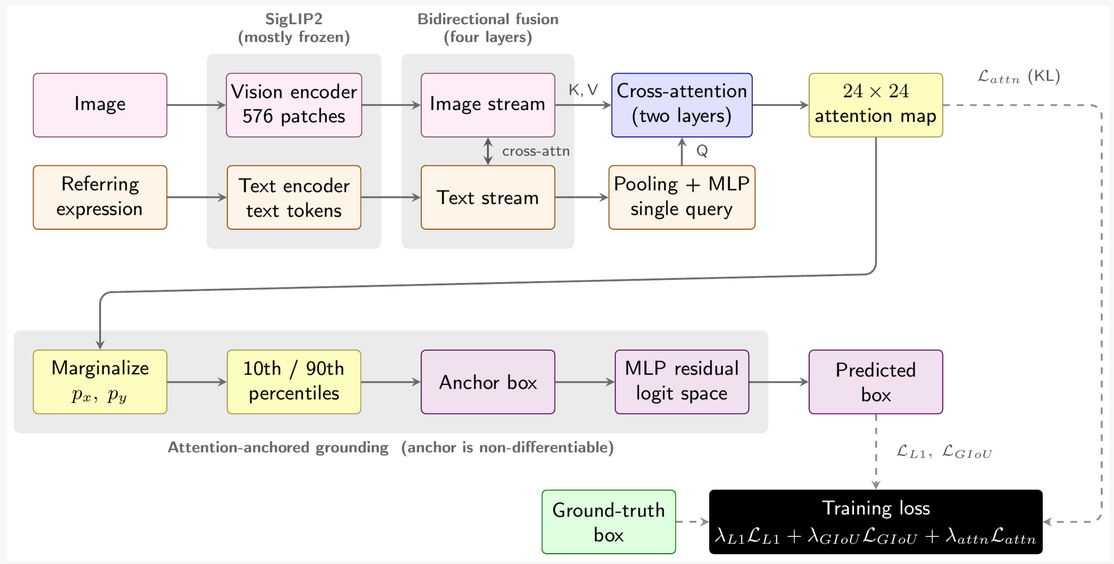
Sources: RefCOCO / RefCOCO+ / RefCOCOg, RefCLEF, FineCops-Ref, FG-OVD, RefDrone, Hindi Visual Genome, RefOI, Multimodal Ground, Toloka VQA and LocateBench, covering everyday scenes, aerial imagery, fine grained attributes, and question / multiple-choice formats. A total of 799,765 unique English expressions are translated to **Chinese, German, Dutch, Spanish, French, Italian, Korean, Portuguese** and **Russian**.

① Translate
The TowerInstruct-7B-v0.2 model translates every English expression into 9 languages.

② Score
The COMETkiwi-DA model performs reference-free quality estimation, and a per-language P_{40} threshold is used to flag weak translations.

③ Enhance
Approximately 3.7M flagged expressions are re-translated with the Gemini 2.0 Flash Lite model, given the image plus the target bounding box.

3 Attention-Anchored Architecture



Frozen SigLIP2 encoders (last 3 layers unfrozen) → 4-layer bidirectional cross-modal fusion → single text query → 2-layer cross-attention decoder → 24×24 attention map → percentile anchor → MLP residual refinement performed in logit space.

From attention to a bounding box

The decoder's cross-attention over the 24×24 patch grid is treated as a 2-D distribution $\mathbf{A}_{2D}$. Marginalizing along each axis and taking the **10th / 90th percentiles** gives a tight, outlier-robust anchor covering the central 80% of the attention mass.

Multi-objective training loss

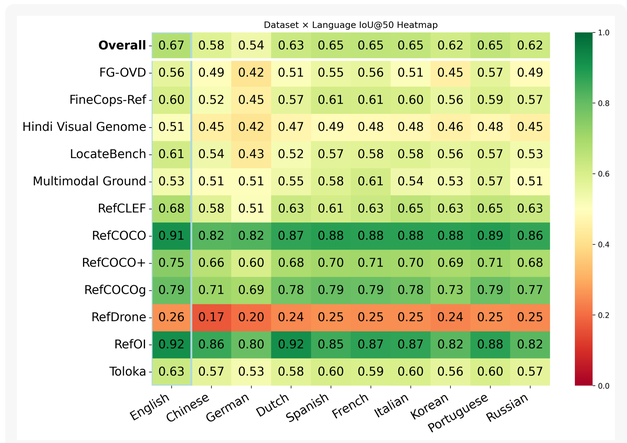
$\mathcal{L} = 5.0\,\mathcal{L}_{L1} + 2.0\,\mathcal{L}_{\text{GIoU}} + 0.2\,\mathcal{L}_{\text{attn}}$, $\quad \mathcal{L}_{\text{attn}} = \mathrm{KL}(p_{\text{target}} \,\|\, a_{\text{norm}})$ vs. the rasterized ground-truth box

4 Comparison with the State of the Art (IoU@50, %)

Dataset	Strongest published baseline	Params	Baseline	Ours (A)	Ours (E)
RefCOCO	CogVLM	17B	**92.0**	86.9	91.3
RefCOCO+	CogVLM	17B	**88.5**	68.9	74.6
RefCOCOg	CogVLM	17B	**89.7**	76.5	79.3
RefCLEF	LMSVA	n/r	**73.0**	61.8	68.1
LocateBench	GPT-4o	n/d	60.4	55.2	**61.4**
RefDrone	NGDINO-T	∼172M	21.8	23.6	**25.7**
FineCops-Ref	CogVLM	17B	**78.2**	56.8	60.3
Toloka VQA	wztxy89	n/r	**83.4**	58.3	62.6
Overall	**12 datasets**			**62.5**	**67.2**

A: aggregate multilingual evaluation over all 10 languages; E: English only. On RefCOCO the model surpasses Grounding DINO (90.8) and OFA-Large (89.3) and approaches the 17B-parameter CogVLM; it exceeds GPT-4o on LocateBench and NGDINO-T on the aerial RefDrone benchmark. Aggregate multilingual mIoU is 0.571 (median 0.721); 50.1% of predictions exceed IoU 0.7.

5 Multilingual Consistency



IoU@50 per dataset (rows) and language (columns).

English sets the performance ceiling at **63.8% IoU@50**. Romance languages cluster within 0.1 pt of each other (Spanish 61.3, French 61.4, Italian 61.3, Portuguese 61.3), only **2.5 pt below English**. Dutch (59.2), Russian (59.1) and Korean (58.7) follow closely; Chinese (55.5) and German (50.8) show the largest gaps.

Multilingual training nearly doubles overall accuracy (31.5% → 62.6% IoU@50, +98.7% relative) and also improves **English** results by +15.1 pt. The languages help each other, pointing to complementarity instead of competition.

6 Ablations & Zero-Shot Baseline (IoU@50, %)

Attention anchoring vs. direct regression

Dataset	Anchored (A/E)	Direct (A/E)	Δ (A/E)
RefCOCO	86.9 / 91.3	79.4 / 87.7	**+7.5** / **+3.6**
RefCOCO+	68.9 / 74.6	57.8 / 62.9	**+11.1** / **+11.7**
RefCOCOg	76.5 / 79.3	65.2 / 69.3	**+11.3** / **+10.0**
RefDrone	23.6 / 25.7	7.5 / 8.8	**+16.1** / **+16.9**
MM Ground	54.4 / 53.2	21.6 / 23.4	**+32.8** / **+29.8**
FG-OVD	51.2 / 56.5	36.9 / 42.4	**+14.3** / **+14.1**
Overall	**62.5 / 67.2**	**54.2 / 59.2**	**+11.3 / +8.0**

Anchoring helps most on small or visually complex targets (Multimodal Ground +32.8, RefDrone +16.1). Variants share the same decoder and differ only in the prediction head.

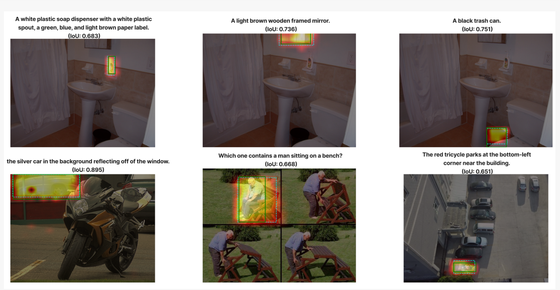
Zero-shot Qwen3-VL-8B vs. ours

Language	Qwen3-VL-8B	Ours (469M)	Δ
English	63.5	**67.2**	**+3.7**
Spanish	62.1	**64.6**	**+2.5**
French	62.7	**64.8**	**+2.1**
Italian	61.7	**64.5**	**+2.8**
Portuguese	61.8	**64.5**	**+2.7**
Dutch	59.9	**62.7**	**+2.8**
Russian	61.9	**62.2**	**+0.3**
Korean	59.5	**61.5**	**+2.0**
Chinese	**59.6**	58.3	−1.3
German	**61.9**	54.1	−7.8
Average	**61.5**	**62.5**	**+1.0**

Qwen3-VL-8B-Instruct is prompted with the image and the native-language expression to output a normalized box, with no task-specific training. Despite ≈17× more parameters, it trails our model overall (61.5 vs. 62.5 IoU@50; mIoU 0.561).

7 Qualitative Examples

"The second wine glass from the left", queried in Russian, English and Italian. Green: ground truth; red: prediction. Across three language families the predicted boxes are nearly identical.



Cross-attention maps that anchor the prediction. Note that the attention concentrates on the referred object, and the percentile anchor plus residual refinement turns it into a tight bounding box.

Small object localization across six languages. "What do cats wear for fashion or identification?" in Russian, Spanish, German, Italian, English, and Portuguese, with IoU between 0.80 and 0.85 in every language.

8 Conclusions

We establish infrastructure for multilingual REC: a unified 10-language dataset of ≈8M referring expressions, and an efficient attention-anchored model that decomposes localization into coarse spatial attention and learned refinement.

Despite its small size, the proposed model competes with far larger English-only systems and with a zero-shot 8B generalist MLLM, while grounding expressions consistently across languages. Efficient multilingual REC that serves users in their native languages is feasible.

Dataset, model weights, and code (dataset construction, training, evaluation) are publicly available:
https://multilingual.franreno.com

Supported by the Portuguese Recovery and Resilience Plan (Center for Responsible AI, C645008882-00000055) and by FCT projects 2024.14126.UTA, 2024.12756.CMU, UID/50021/2025, UID/PRR/50021/2025, LA/P/0083/2020, UIDP/50009/2020 and UIDB/50009/2020.