Blind to Position, Biased in Language: Probing Mid-Layer Representational Bias in Vision-Language Encoders for Zero-Shot Language-Grounded Spatial Understanding

Na Min An[1*], Inha Kang[1*], Minhyun Lee[2], and Hyunjung Shim[1†]

[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]Samsung Electronics

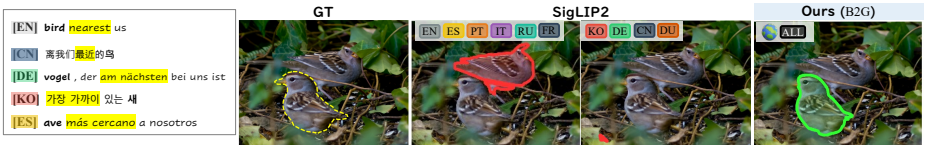
Abstract. Vision–Language Encoders (VLEs) are widely adopted as the backbone of zero-shot referring image segmentation (RIS), enabling text-guided localization without task-specific training. However, prior works underexplored the underlying biases within mid-layer representations that preserve positional and language-specific information. Through layer-wise investigation, we reveal that the conventionally used final-layer multimodal embeddings prioritize global semantic alignment, leading to two coupled consequences. First, vision embeddings exhibit weak sensitivity to positional cues. Second, multilingual text embeddings form language-dependent geometric shifts within the shared space. Motivated by these findings, we identify an underexplored pathway within VLE mid-layers to construct a spatial map, applicable for improving zero-shot RIS by 1–7 mIoU on nine RefCOCO benchmarks. Furthermore, leveraging mixed-language mid-layer embeddings yields enhanced spatial grounding accuracy (+7–8 mIoU and IoU@50), albeit with increased inference cost, and also improves performance on the zero-shot text-to-image retrieval task. Our work opens up the discussion about the effects of effective representational bias probing of VLEs for enhanced spatial grounding. The code is available at https://github.com/kaist-cvml/Biased2Grounded.

Keywords: Spatial Bias · Multilingual Bias · Vision-Language Encoder

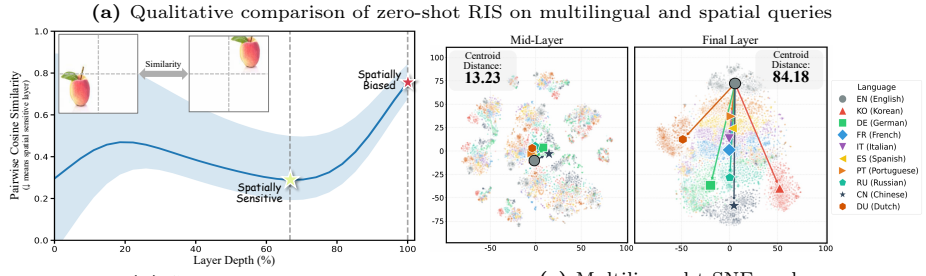
1 Introduction

Language-grounded spatial understanding is a fundamental capability for embodied AI and interactive vision. Tasks such as language-conditioned navigation and manipulation require agents to map free-form instructions to precise regions and actions within cluttered scenes [1,4,25,28,29]. In real deployments, the same intent may be expressed through diverse phrasings and across multiple languages. This motivates multilingual grounding benchmarks that systematically extend English datasets via translation, reaching a million-expression scale [11,23]. Despite this

*Equal contribution. Contact: naminan@kaist.ac.kr, rkswlsj13@kaist.ac.kr
†Corresponding author.



(a) Qualitative comparison of zero-shot RIS on multilingual and spatial queries



(b) Spatial probe **(c)** Multilingual t-SNE probe

Fig. 1: Task and preliminary diagnosis. (a) Our proposed B2G accurately localizes referred regions in multilingual zero-shot RIS. **(b)** A tile-permutation probe reveals intermediate layers of vision encoders are spatially sensitive, while the final embedding is nearly position-invariant. **(c)** Semantically equivalent multilingual queries exhibit systematic geometric shifts in the latent space for the final layer compared to the middle layer.

progress, recent analyses still reveal persistent weaknesses in fine-grained spatial reasoning [9, 36] as well as language-dependent representational biases [48].

These two issues stem from a shared architectural bottleneck: Most systems rely on the final-layer global embedding of pretrained Vision-Language Encoders (VLEs) for cross-modal scoring. Common multimodal learning approaches [26, 37] concentrate alignment on this single representation, causing diverse visual and linguistic signals to collapse into a compact semantic space. As deeper layers increasingly optimize for high-level abstraction, they progressively discard fine-grained positional structure while introducing language-dependent geometric shifts, likely influenced by the training imbalance and tokenization asymmetry. Crucially, precise spatial cues and language-stable representations remain encoded within the *intermediate* layers, prior to the final abstraction bottleneck.

The limitation becomes particularly evident in referring image segmentation (RIS), which evaluates whether a VLE can ground language to spatially precise image regions. RIS therefore provides a natural testbed for diagnosing these two biases, as the task isolates the core cross-modal challenge of mapping language descriptions to pixel-level masks. In particular, *zero-shot* RIS strongly penalizes spatially insensitive or cross-lingually unstable scoring interfaces, especially under translation-based multilingual benchmarks where the image-mask pair remains fixed while only the query language varies.

Pretrained VLEs [6, 14, 26, 37, 47] have now become standard backbones for zero-shot RIS, enabling mask grounding without any dataset-specific fine-tuning [7, 18, 19, 22, 32, 34, 39, 45] (Fig. 1a). Most training-free pipelines follow a proposal-and-rank design: an off-the-shelf segmentor generates candidate masks [5, 10, 16, 38], and a frozen VLE ranks them using image–text similarity. Since modern segmentors

often produce multiple overlapping or plausible masks, successful zero-shot RIS critically depends on whether the VLE's ability to disambiguate the query and localize the target region. When the ranking signal is dominated by global semantics, models frequently exhibit "opposite visualization" [15] (visualized in Fig. 3 and further analyzed in the Appendix), selecting background context instead of the intended foreground target. This behavior highlights that the final-layer embedding is a suboptimal interface for pixel-level grounding. Yet existing training-free improvements mainly address spatial disambiguation under English queries, while multilingual robustness typically relies on training new encoders—leaving no unified training-free solution that simultaneously mitigates both spatial and language biases on frozen VLEs.

Our analysis pinpoints where the missing cues remain accessible within the model. To identify position-sensitive visual representations, we introduce a spatial equivariance probe based on controlled image tile permutation. Intermediate layers become increasingly sensitive to spatial arrangement, whereas the final embedding is nearly position-invariant (Fig. 1b). We observe an analogous phenomenon in multilingual text embeddings. Even for semantically identical sentences, the final layers exhibit systematic language-dependent shifts (Fig. 1c). While these shifts show little impact on text-to-text similarity, they significantly distort cross-modal similarity with visual features. Together, spatial and multilingual cues remain encoded within pretrained VLEs but are attenuated at the final alignment layer, and can be recovered by selectively leveraging intermediate representations at test time.

Building on this insight, we propose **B**iased **to** **G**rounded (B2G), a training-free framework following a unified *probe* → *select* → *ground* paradigm. Rather than replacing existing pipelines, B2G acts as a plug-in that complements the globally aligned but spatially/linguistically compressed final-layer representations with mid-layer spatial/linguistic bias cues. For spatial grounding, B2G extracts features from a position-sensitive intermediate vision layer to construct a dense patch-level similarity map (`P-Map`), whose high-response regions form spatial clusters that guide mask reranking. To improve multilingual robustness, B2G probes the text encoder across N translations to identify a language-stable layer and compute a mid-layer multilingual centroid, which is injected into the anchor query to stabilize representation geometry before final contextualization. The framework operates entirely on frozen VLEs, supporting both fixed (B2G_{fix}) and automatic ($\text{B2G}_{\text{dyn-clus}}$, $\text{B2G}_{\text{dyn-perm}}$) layer selection, without fine-tuning, gradient backpropagation [27], or recomputation [2, 15].

B2G consistently improves zero-shot RIS across standard benchmarks [20, 21, 41], surpassing the strongest prior training-free baseline by 2.15 mIoU on RefCOCO, 5.47 mIoU on RefCOCO+, and 5.75 mIoU on RefCOCOg (CLIP ViT-B/16) [18]. Ours also substantially improves multilingual zero-shot RIS, increasing average IoU@50 from 45.13 to 53.41 and mIoU from 42.42 to 49.62 across nine translated benchmarks [23]. Beyond RIS, our mid-layer multilingual stabilization also improves robustness in zero-shot multilingual text-to-image retrieval (+4.22, +2.77, +6.42 Recall@1 on RefCOCO, RefCOCO+, RefCOCOg).

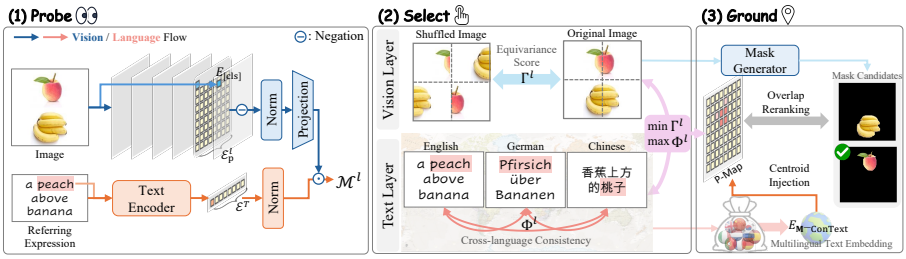


Fig. 2: Overview of Proposed Framework. B2G follows a unified ***probe*** → ***select*** → ***ground*** paradigm. It *probes* mid-layer VLE representations to identify spatially sensitive and language-stable embeddings, *selects* the most informative layers, and performs mask *grounding* through spatial map, `P-Map`, with multilingual centroid fusion.

These results suggest that spatial and multilingual biases largely stem from the final-layer abstraction interface rather than the absence of grounding information itself. By probing intermediate representations, B2G unlocks spatially precise and language-robust grounding from frozen VLEs.

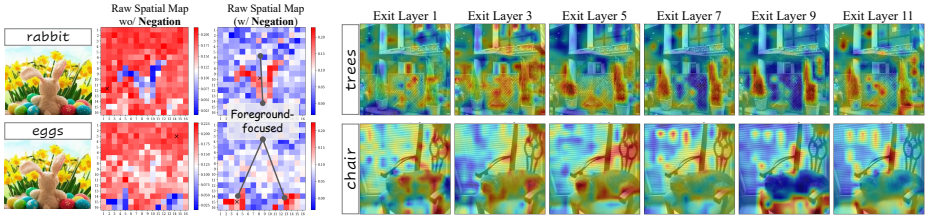
2 Methodology

Problem setup. In zero-shot RIS, we are given an image I, a referring expression T, and K mask proposals $\{M_k\}_{k=1}^{K}$ from an off-the-shelf segmentor. Let $\mathcal{R}(I, M_k)$ denote the region extraction operator used by prior training-free pipelines. Following standard practice, we obtain a region embedding (of masked or cropped images specified by M_k) by applying the frozen vision encoder to $\mathcal{R}(I, M_k)$ and pooling its final-layer representation. We denote by $E_{\text{txt}}(T)$ the text embedding used for scoring (Sec. 2.2 replaces it with a multilingual-stabilized variant). A frozen VLE then scores each proposal by global image–text similarity:

$$S_k = \cos\Big(E_{\text{vis}}(\mathcal{R}(I, M_k)),\ E_{\text{txt}}(T)\Big), \qquad \hat{M} = \arg\max_k S_k. \tag{1}$$

B2G keeps the proposal generator and $\mathcal{R}(\cdot)$ unchanged, and focuses on strengthening the grounding signal by selecting more informative internal representations and adding a lightweight reranking cue on top of the baseline score.

Proposed framework. Our key premise is that the final-layer alignment interface of pretrained VLEs is optimized for global semantic matching, which can attenuate the structural cues required for precise grounding and further amplify the English-dominant alignment bias introduced during pretraining. This manifests as (1) weakened positional sensitivity in vision features and (2) language-dependent geometric shifts in multilingual text features (Fig. 1b–c). B2G therefore follows a unified *probe* → *select* → *ground* paradigm (Fig. 2): (1) *probe* mid-layer representations to identify structure-preserving layers. (2) *select* a position-sensitive vision layer and a language-stable text layer. (3) perform a mask *ground*ing using patch-level similarity and multilingual centroid injection.



(a) Raw patch-level spatial map **(b)** Layer-wise final spatial map (`P-Map`)

Fig. 3: Spatial Map Visualization. (a) Raw spatial map with and without negation operation. **(b)** Final spatial maps generated across varying exit layers.

2.1 Biased to Grounded (B2G) for (Non-Multilingual) VLE

Stage 1: Spatial bias *probing*. The central challenge for spatial grounding at intermediate layers is that patch-level embeddings tend to suppress foreground activations in favor of global context, pushing target-region similarities into the negative domain [15]. Our approach addresses this via two steps. We first construct a raw patch–text similarity map to localize spatial responses. We then apply a deterministic sign inversion to correct the foreground suppression. This is because the referred foreground patches become systematically anti-correlated with the query direction in the joint embedding space at several mid-layer exits, leading to inverted cosine responses (Fig. 3a). Since $\cos(-u, v) = -\cos(u, v)$, a deterministic sign flip converts this inverted activation into an intuitive, target-high map. We apply the flip to patch features before the shared projection[1], which is empirically stable across VLE backbones.

Specifically, given an input image I, we extract patch-level embeddings from the l–th exit layer of the visual encoder, $\mathcal{E}_{\mathrm{p}}^{l} \in \mathbb{R}^{p\times p\times d^{*}}$. To compare them with text in the joint embedding space, we use the model's post LN (per block) and project them using the pretrained projection matrix $\mathbb{W}_{d^{*}}^{d}$:

$$E_{\mathrm{p}}^{l} = \mathrm{LN}(\mathcal{E}_{\mathrm{p}}^{l}) \cdot \mathbb{W}_{d^{*}}^{d} \in \mathbb{R}^{p\times p\times d}.$$

For the text embedding, we use E_{ConText}, our hybrid text representation that fuses global sentence-level features with context-augmented local noun features (`CT`, Appendix for details) for scoring; Sec. 2.2 replaces it with the multilingual-stabilized $E_{\mathbf{M}\text{-}\mathrm{ConText}}$. We compute a raw patch-level similarity map as follows:

$$\widetilde{\mathcal{M}}^{l} = (\widetilde{\mathcal{M}}_{ij}^{l})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}, \quad \widetilde{\mathcal{M}}_{ij}^{l} = \cos\big(E_{\mathrm{p}}^{l}[i,j],\, E_{\mathrm{ConText}}\big)\,.$$

However, direct cosine similarity at mid layers may exhibit the known "opposite visualization" phenomenon [15], where foreground patches receive relatively low responses. To deterministically correct this effect, we negate the patch embeddings before normalization, yielding our raw spatial map (later interpolated to yield

[1]For models that do not have an image-to-text projection layer, such as SigLIPs, we use a post-layer normalization (LN) layer instead.

the final spatial map, `P-Map`), $\mathcal{M}^l \in \mathbb{R}^{p\times p}$:

$$\widetilde{E}^l_{\mathrm{p}}[i,j] = \mathrm{LN}\big(-\mathcal{E}^l_{\mathrm{p}}[i,j]\big)\cdot \mathbb{W}^d_{d^*}, \qquad \mathcal{M}^l_{ij} = \cos\Big(\widetilde{E}^l_{\mathrm{p}}[i,j],\, E_{\mathrm{ConText}}\Big).$$

This sign inversion restores foreground-localized activations and produces more interpretable spatial maps without additional training time and memory costs.

Stage 2: Vision layer *selection*. Although intermediate layers exhibit varying degrees of spatial sensitivity, not all layers yield a valid spatial map (Fig. 3). We therefore select a position-sensitive layer using a tile-permutation probe. Given an input image I, we construct a spatially permuted variant I^π by shuffling non-overlapping tiles over the $p \times p$ grid. This preserves local appearance statistics while disrupting global spatial configuration.

For each candidate layer l, we extract patch embeddings from both I and I^π, project them as above, and restore spatial correspondence by applying the inverse permutation π^{-1} to the shuffled patch features. We then compute an equivariance score as follows:

$$\hat{E}^l_{\mathrm{p}}(I^\pi) = \pi^{-1}\big(E^l_{\mathrm{p}}(I^\pi)\big), \qquad \mathit{\Gamma}^l = \frac{1}{p^2}\sum_{i=1}^{p^2}\cos\Big(E^l_{\mathrm{p}}(I)[i],\, \hat{E}^l_{\mathrm{p}}(I^\pi)[i]\Big).$$

Larger $\mathit{\Gamma}^l$ indicates that patch features remain similar after permutation (*i.e.*, weak dependence on spatial arrangement), while smaller values indicate stronger position sensitivity.

To determine the optimal layer (l^*_{vis}), we investigate three selection strategies using a calibration set: dynamic selection via a permutation probe ($\mathrm{B2G}_{\mathrm{dyn\text{-}perm}}$), cluster coherence ($\mathrm{B2G}_{\mathrm{dyn\text{-}clus}}$), and a fixed mid-layer heuristic ($\mathrm{B2G}_{\mathrm{fix}}$). Empirically, all three strategies yield highly comparable grounding performance (details in the Appendix). This consistency suggests that spatially informative structure is a recurring mid-layer tendency across VLEs, although its map sharpness and optimal layer remain backbone-dependent rather than universally fixed. This motivates our dynamic variants, especially for backbones whose mid-layer maps are more diffuse. Consequently, B2G can reliably operate with the minimal cost $\mathrm{B2G}_{\mathrm{fix}}$ strategy in practice without sacrificing accuracy (detailed formulations are provided in the Appendix).

Stage 3: Mask *grounding* by overlap reranking. Given `P-Map` $\mathcal{M}^{l^*_{\mathrm{vis}}} \in \mathbb{R}^{p\times p}$, we convert it into a spatial guidance mask. We normalize $\mathcal{M}^{l^*_{\mathrm{vis}}}$ to $[0,1]$, threshold it with δ, and extract connected components on the $p \times p$ grid to obtain clusters $\{C_r\}$. Each cluster is upsampled to the image resolution, and we merge them into a binary guidance mask G.

For each proposal M_k, we compute the baseline similarity score S_k defined in Eq. (1), alongside an overlap score O_k. We use $O_k = \mathrm{IoU}(M_k, G)$ by default; using the mean `P-Map` response within M_k yields similar trends (Appendix). We subsequently rerank the proposals using a combined score:

$$S'_k = S_k + \lambda\, O_k, \qquad \hat{M} = \arg\max_k S'_k. \tag{2}$$

To reduce noise, overlap reranking is applied only to the top-κ proposals under S_k. Overall, B2G extracts dense spatial cues from frozen mid-layer features and uses them as a lightweight reranking signal (denoted as `TC`; Top Candidate selection), requiring neither gradient-based saliency nor recomputation of attention.

B2G shows that the *final-layer VLE interface* is spatially weak, rather than indicating an absence of spatial knowledge in the VLE itself, by revealing grounding-relevant cues from intermediate representations. Unlike prior methods that rely on computationally intensive gradient-based saliency or attention maps, B2G derives spatial guidance from a simple yet effective spatial map, `P-Map`. This training-free extraction avoids recomputation and remains faithful to the model's intrinsic spatial structure.

Our spatial probe demonstrates that VLE mid-layers preserve position-sensitive structures that are largely attenuated at the final layer. Interestingly, a similar phenomenon emerges in the text encoder, where intermediate text layers maintain a more language-consistent embedding geometry compared to the final representation. This structural parallel motivates our unified framework, which selects the most informative mid-layers from both modalities as detailed in the subsequent section.

2.2 Biased to Grounded (B2G) for Multilingual VLE

Stages 1–2: Multilingual bias *probing* and layer *selection*. When the same referring expression is translated into multiple languages, the resulting text embeddings should ideally lie in the same region of the shared multimodal space. In practice, however, language-specific surface variations, such as morphology, word order, and tokenization, introduce systematic geometric shifts that accumulate across transformer layers.

Given an input expression T, we first generate semantically aligned translations $\{T^{(1)}, \ldots, T^{(N)}\}$ across N languages, designating the initial expression $T^{(1)} = T$ as the anchor language. We then process these expressions through the text encoder up to layer l. From this layer, we extract a pooled sentence representation $h^l_{(n)} \in \mathbb{R}^d$ utilizing the hidden state of the end-of-text token. Subsequently, we quantify the cross-language consistency by

$$\Phi^l = \frac{1}{\binom{N}{2}} \sum_{a<b} \cos\Big(h^l_{(a)}, h^l_{(b)}\Big) .$$

Larger Φ^l indicates a more language-stable geometry at layer l. We select the most consistent layer by $l^*_{\text{txt}} = \arg\max_l \Phi^l$.

Stage 3: Multilingual centroid for *grounding*. To stabilize the language-dependent geometry, we form a multilingual centroid at the selected intermediate layer l^*_{txt}. Given translations $\{T^{(n)}\}_{n=1}^N$, we first obtain their token-level embeddings and propagate them through the text encoder to extract pooled mid-layer embeddings $h^{l^*_{\text{txt}}}_{(n)}$ and compute their average $\bar{h}^{\,l^*_{\text{txt}}}$ across languages. We then inject

this centroid into the anchor-language hidden state at layer l^*_{txt} by replacing the end-of-text (EOT) token representation. Let $\mathcal{H}_{(n)}^{l^*_{\text{txt}}}$ denote the token-wise hidden states of $T^{(n)}$ at layer l^*_{txt} and *eot* its EOT index. We set $\mathcal{H}_{(n)}^{l^*_{\text{txt}}}[eot] \leftarrow \bar{h}^{\,l^*_{\text{txt}}}$ and also integrate the centroid of the global sentence context and local noun-phrase context. $\mathcal{E}_{\mathbf{M}\text{-ConText}} = \gamma\mathcal{E}_{\mathbf{M}\text{-Sen}} + (1-\gamma)\mathcal{E}_{\mathbf{M}\text{-NounCtx}}$. Then, this merged embedding is propagated through the subsequent layers and the final projection head to obtain the final embedding $E_{\mathbf{M}\text{-ConText}}$. By performing both multilingual and contextual fusion at an intermediate layer, the model preserves language-invariant semantic structure before later layers amplify language-specific variations, producing a stable language-agnostic feature while keeping the text encoder frozen.

Test-time usage and mask scoring. During inference, B2G pipeline operates without gradient updates. First, the text query $E_{\mathbf{M}\text{-ConText}}$ is computed strictly once, restricting the multilingual overhead to N early-exit partial passes and a single continuation pass. Second, for an image, a single forward pass through the frozen vision encoder extracts mid-layer patch embeddings. These are combined with $E_{\mathbf{M}\text{-ConText}}$ to construct the `P-Map` and generate the spatial guidance mask G. Finally, we evaluate the similarity S_k (Eq. (1)) between $E_{\mathbf{M}\text{-ConText}}$ and all κ candidates. To resolve spatial ambiguities, we rerank the top-κ candidates by combining S_k with their overlap O_k against G to produce the final mask $\hat{M}$.

3 Experiments

3.1 Experimental Setup

Tasks and evaluation protocols. We evaluate B2G under three complementary protocols: standard English zero-shot RIS, multilingual zero-shot RIS, and zero-shot multilingual text-to-image retrieval. Unless otherwise noted, B2G is used as a training-free plug-in on frozen VLEs under the same pipeline.

Evaluation datasets and metrics. Following prior zero-shot RIS works [18, 22, 33, 34, 39, 45], we report standard RIS results on RefCOCO [21], RefCOCO+ [21], and RefCOCOg [20], which differ in linguistic complexity and positional cues– RefCOCO contains more explicit spatial terms, while RefCOCOg features longer expressions. We also report results on PhraseCut [41] under *all* and *unseen* (COCO class [17]) settings. For RIS, the main metrics are mean Intersection over Union (mIoU) (if unspecified), overall IoU (oIoU), and IoU@50, which measures the percentage of predictions whose IoU with the ground-truth mask exceeds 0.5.

Implementation details. To ensure a fair comparison with previous zero-shot RIS methods, we use the same frozen VLE backbones as the compared baselines, including CLIP (ViT-B/32 and ViT-B/16) [26], as well as BLIP (Appendix) [14], SigLIP [47], DFN [6], and SigLIP2 [37]. Since B2G_{fix} uses a fixed exit layer, we determine the optimal exit layer (l^*) of the visual encoder and the initial threshold (δ) on a small unlabeled calibration split comprising 10% of the RefCOCOg val (U) set, and keep them fixed at test time.

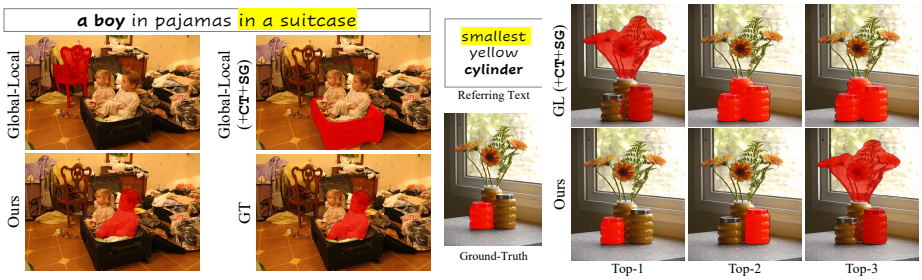
Table 1: Comparison of Zero-shot RIS Performance. Ours consistently achieves superior performance over existing feature extraction approaches, Global-Local [45] and HybridGL [18], across different backbone encoders. For fair comparison, all the listed methods have been applied with `CT` and `SG` (ablation in later). The baseline is in red, and statistically significant improvements over the baseline are in blue for each backbone.

Method	Pre-trained Segmentor	RefCOCOg val	RefCOCOg test	RefCOCOg valG	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	Avg.
CLIP ViT-B/32											
Global-Local	SAM	44.71	45.97	45.75	35.99	35.36	36.15	32.50	33.62	30.56	37.85
HybridGL	SAM	46.54	47.00	47.34	40.67	41.84	39.05	37.52	40.73	33.40	41.57
+B2G	SAM	46.89	47.20	47.67	41.07	42.15	39.41	38.54	42.12	33.89	42.55 (+0.98)
Global-Local	Mask2 Former	47.44	46.78	47.68	43.19	46.81	39.32	39.47	44.74	34.19	43.29
HybridGL	Mask2 Former	48.83	48.35	49.26	41.72	44.46	37.64	40.91	47.11	33.28	43.51
+B2G$_{\text{fix}}$	Mask2 Former	53.47	53.84	53.88	48.58	54.39	42.01	46.34	53.60	36.80	49.21 (+5.70)
+B2G$_{\text{dyn-clus}}$	Mask2 Former	53.56	53.62	53.83	48.62	54.37	41.57	45.73	52.57	36.36	48.91 (+5.40)
+B2G$_{\text{dyn-perm}}$	Mask2 Former	53.42	53.76	54.07	48.55	54.22	41.68	46.07	53.22	36.28	49.03 (+5.52)
CLIP ViT-B/16											
Global-Local	SAM	48.54	49.49	48.82	40.06	40.00	40.36	35.15	37.33	33.35	41.90
HybridGL	SAM	50.37	50.52	50.41	44.74	46.48	43.26	40.17	44.44	36.19	45.18
+B2G	SAM	50.75	50.74	50.77	45.17	46.82	43.65	41.28	45.94	36.72	46.24 (+1.06)
HybridGL	Mask2 Former	49.53	50.05	49.92	45.36	49.44	40.78	41.81	47.60	35.34	45.54
Global-Local	Mask2 Former	49.29	49.26	48.96	47.38	53.11	41.49	40.86	45.96	35.34	45.74
+B2G$_{\text{fix}}$	Mask2 Former	54.42	55.62	54.71	50.05	55.38	43.01	46.83	54.08	37.67	50.20 (+4.46)
+B2G$_{\text{dyn-clus}}$	Mask2 Former	54.37	55.23	54.58	49.97	55.04	43.48	46.21	52.50	37.61	49.89 (+4.15)
+B2G$_{\text{dyn-perm}}$	Mask2 Former	54.57	55.45	54.63	50.18	55.47	43.71	46.64	52.89	37.63	50.13 (+4.39)
SigLIP ViT-B/16											
Global-Local	Mask2 Former	49.67	49.33	49.87	48.77	55.00	41.39	42.64	48.98	33.97	46.62
+B2G$_{\text{fix}}$	Mask2 Former	52.35	51.80	52.01	46.05	51.34	39.49	46.10	52.66	37.47	47.70 (+1.08)
+B2G$_{\text{dyn-clus}}$	Mask2 Former	52.43	51.90	51.97	46.67	52.42	40.78	45.73	52.22	37.74	47.98 (+1.36)
+B2G$_{\text{dyn-perm}}$	Mask2 Former	52.12	51.43	51.78	45.56	51.29	39.72	45.95	52.36	37.66	47.54 (+0.92)
DFN VIT-H/14											
Global-Local	Mask2 Former	48.84	49.19	48.84	45.74	52.85	38.53	39.11	46.17	32.51	44.64
+B2G$_{\text{fix}}$	Mask2 Former	56.23	55.88	55.68	50.91	56.59	44.79	47.05	54.12	38.91	51.13 (+6.49)
+B2G$_{\text{dyn-clus}}$	Mask2 Former	55.74	55.98	55.57	51.26	57.30	44.42	47.38	54.08	38.66	51.60 (+6.96)
+B2G$_{\text{dyn-perm}}$	Mask2 Former	55.35	55.94	55.26	51.10	57.44	43.60	47.25	54.39	38.19	50.95 (+6.31)

3.2 Main Results on Standard Zero-shot RIS

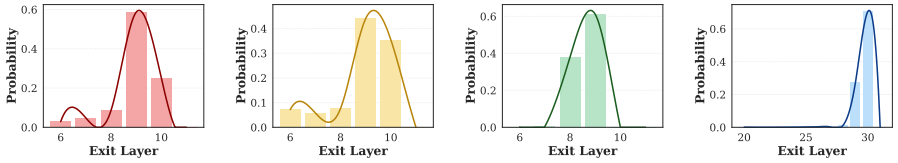
Zero-shot RIS. As shown in Tab. 1 and Fig. 4 (oIoU and different VLE results in the Appendix), B2G consistently yields additional improvements, even when integrated with superior feature extraction methods [18,45] enhanced by both `CT` and `SG` (spatial guider implemented from [18]) across varying VLE backbones[2]. Particularly, B2G attains noticeable gains for DFN [6], where it achieves +6.49, +6.96, and +6.32 mIoU with B2G$_{\text{fix}}$, B2G$_{\text{dyn-clus}}$, and B2G$_{\text{dyn-perm}}$, respectively. Across CLIP ViT-B/32 and CLIP ViT-B/16 (SAM), the performance differences among selection variants are negligible; for readability, we denote them uniformly as B2G. Mask2Former shows better synergy with our method, similar to Wang *et al.* [39], due to its ability to generate less noisy, semantic masks [5,16]. Fig. 4b

[2]Note that the improvement of SigLIP is smaller than the other models due to the quality of `P-Map` (Appendix).



(a) Comparison of Top-1 Selected Masks **(b)** Comparison of Top-κ Candidate Masks

Fig. 4: Qualitative Comparison on Zero-shot RIS. (a) B2G shows strong spatial grounding ability. **(b)** The masks generated using `P-Map` are diverse and accurate.



(a) CLIP ViT-B/32 **(b)** CLIP ViT-B/16 **(c)** SigLIP ViT-B/16 **(d)** DFN ViT-H/14

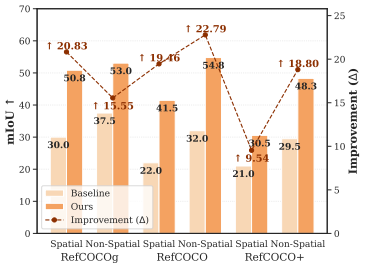
Fig. 5: Automatically Selected Layers. We observe a high probability concentration in middle layers automatically selected for $\text{B2G}_{\text{dyn-perm}}$ ($n = 66\text{k}$).

indicates that the top candidate masks using our `P-Map` are highly informative in spatial grounding, as they yield more diverse top-κ candidate masks than the previous method.

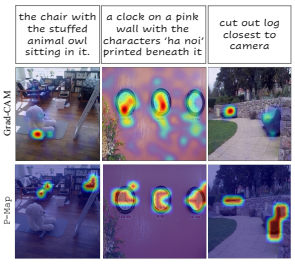
3.3 Spatial Grounding Analysis and Challenging Benchmarks

Selected intermediate layers. Different VLEs exhibit distinct layer-wise bias profiles, possibly stemming from variations in their training objectives, projection heads, and tokenization. Consequently, the layer that best preserves spatial equivariance or multilingual consistency is not universal across backbones. This motivates our automatic layer selection strategy: rather than prescribing a fixed layer for each model family, B2G explicitly identifies where structural properties are preserved and adaptively selects the most informative intermediate layer. As shown in Fig. 5, the selected layers mostly concentrate in mid-depth regions across diverse backbones. While the optimal layer may vary, this consistent mid-depth concentration supports our claim that intermediate representations preserve grounding-relevant structure, offering a reliable localization signal.

Spatial grounding analysis. To evaluate the fine-grained spatial grounding ability of VLE (CLIP ViT-B/16), we conduct a targeted evaluation on samples that contain explicit spatial cues (*e.g.*, "left", "closest to"). To identify such samples, we utilize Qwen2.5-14B-Instruct [42] to classify each referring expression as spatial or non-spatial and extract the cue if present (prompt details in the



(a) Subset performance

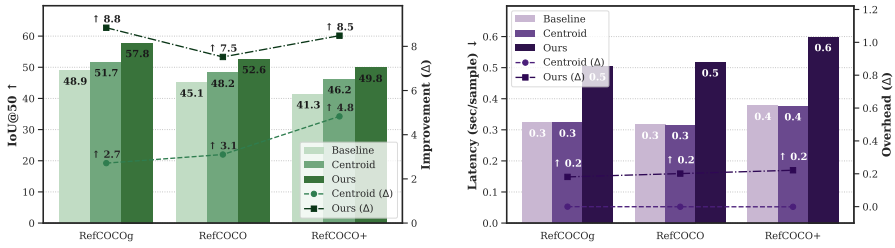


(b) Spatial map comparison

Method	Train?	All	Unseen
CRIS [40]	✓	16.30	14.62
GL† [45]	✗	22.43	21.54
TAS [34]	✗	25.64	24.66
LAVT [44]	✓	28.68	29.14
PRIS [46]	✓	32.75	33.52
IRP [39]	✗	38.10	37.90
HGL† [18]	✗	37.60	37.91
B2G	✗	**38.53**	**38.31**

(c) PhraseCut results

Fig. 6: Challenging Dataset Performance. Ours achieves higher and more accurate performance in both spatial and non-spatial subsets and the PhraseCut dataset.



(a) Spatial grounding performance **(b)** Inference cost

Fig. 7: Comparison of Zero-shot Multilingual RIS and Inference-Cost Performance. The proposed mid-layer representation extraction method achieves **(a)** higher spatial grounding performance (+8–9%) **(b)** at increased inference cost (×1.5–1.7).

Appendix). These automatically generated annotations take up 50.72%, 62.40%, and 23.48% of the validation samples in RefCOCO benchmarks. Fig. 6a shows significant gains using B2G over the baseline, with +20.83% and +15.55% mIoU improvements on the spatial and non-spatial subsets of RefCOCOg.

These improvements demonstrate that our spatial map `P-Map` contributes to the gains on both spatial and non-spatial subsets with accurate localization (Fig. 3). For instance, we observe that `P-Map` better captures target objects in many cases (*e.g.*, "chair" in the first column) than Grad-CAM [27] (Fig. 6b). Comprehensively, `P-Map` achieves better average performance (*e.g.*, +6.9 mIoU) with lower inference latency (*e.g.*, −2.95 *sec.* per sample, details in the Appendix). This leads to our approach achieving the best results even on the very challenging PhraseCut dataset. We obtain oIoU scores of 38.53 (*all*) and 38.31 (*unseen*), outperforming SoTA zero-shot and supervised methods (Fig. 6c). These results highlight the robustness and generalizability of B2G across various datasets.

3.4 Multilingual Grounding Results

Zero-shot multilingual RIS. Figs. 7 and 8 demonstrate that the proposed mid-layer multilingual centroid fusion strategy consistently improves spatial grounding

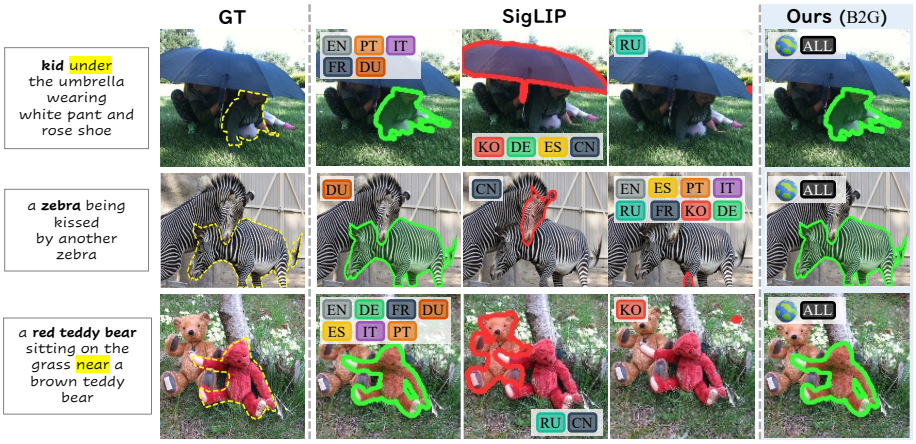


Fig. 8: Qualitative on Zero-shot Multilingual RIS. B2G leverages the centroid of mid-layer multilingual embeddings for robust and language-consistent spatial grounding.

results over both the SigLIP2 baseline and the final-layer centroid fusion strategy. The baseline achieves an average of 45.13 IoU@50 across all nine benchmarks, which increases to 48.67 with centroid aggregation and further to 53.41 with the proposed B2G. A similar trend is observed for mIoU: 42.42 → 46.07 (+3.65) → 49.62 (+7.20). These improvements incur modest additional computation from mid-layer feature extraction and similarity computation, reflecting an unavoidable accuracy–cost trade-off (Fig. 7b).

Qualitative multilingual grounding. Figure 8 illustrates the efficacy of B2G in mitigating both spatial blindness and language-induced biases. The baseline frequently exhibits severe localization failures due to geometric drift when processing non-English queries and struggles with explicit spatial relations (*e.g.*, confusing the actor and receiver in *a zebra being kissed*). By probing structure-preserving intermediate layers, B2G constructs a `P-Map` to recover precise positional cues. Concurrently, injecting a language-stable multilingual centroid at the selected text layer alleviates language-dependent geometric shifts across translated expressions. This stabilization yields a more language-consistent representation and improves grounding across diverse languages without parameter updates, and even corrects baseline failures on standard English queries. Note that text-side B2G increases 064 latency by 1.5×–1.7× with +0.01G memory.

3.5 Zero-shot Multilingual Text-to-Image Retrieval

To validate our method's language-agnostic semantics, we evaluate zero-shot multilingual text-to-image retrieval. As shown in Tab. 2, the baseline final-layer embeddings of SigLIP2 suffer from severe geometric drift in non-English languages. While English achieves a robust 49.44% average Recall@1, languages

such as German and Chinese collapse to 10.16% and 35.53%, respectively. Without any parameter updates, our mid-layer centroid injection yields substantial improvements across all nine non-English languages, achieving average gains of 10.39% for DE and 5.28% for CN. B2G avoids typical generalization trade-offs by not only rescuing underperforming languages but also further enhancing the strong English baseline (+1.23% average gain). This confirms that our method acts as a targeted structural regularizer, effectively unlocking equitable and high-performance multilingual retrieval from biased VLEs.

3.6 Ablation and Sensitivity Studies

Component ablation. Tab. 3a first highlights the individual and combined effects of E_{ConText} (`CT`), top-candidate (`TC`) mask selection via `P-Map`, and the spatial guider (`SG`). Incorporating `CT` and `TC`–both novel components of our framework–consistently improves VLE performance, with particularly notable gains on CLIP ViT-B/32 (+2.46 and +1.32 mIoU and oIoU, respectively). Importantly, the improvements increase with each component added, indicating that leveraging mid-layer representations is critical to overall performance gains.

Table 2: Zero-Shot Multilingual Retrieval Results (Recall@1). We compare the SigLIP2 and B2G. Across all datasets and 10 diverse languages, particularly rescuing languages that suffer from severe geometric drift without degrading the other languages.

Dataset	Method	Target Language (Recall@1 ↑)										Avg.
		EN	DE	FR	IT	ES	PT	RU	KO	CN	DU	
RefCOCO	SigLIP2	61.37	13.67	52.42	45.53	49.29	53.94	52.45	46.83	45.19	45.87	46.66
	+B2G	**62.67** +1.30	**25.82** +12.15	**55.56** +3.14	**47.83** +2.30	**50.44** +1.15	**56.40** +2.46	**56.37** +3.92	**52.00** +5.17	**51.41** +6.22	**50.31** +4.44	**50.88** +4.22
RefCOCO+	SigLIP2	30.96	3.96	19.34	18.52	21.05	21.05	20.76	20.80	21.84	19.82	19.81
	+B2G	**31.64** +0.68	**9.43** +5.47	**21.49** +2.15	**22.03** +3.51	**23.95** +2.90	**23.50** +2.45	**23.37** +2.61	**23.77** +2.97	**24.23** +2.39	**22.42** +2.60	**22.58** +2.77
RefCOCOg	SigLIP2	56.00	12.85	48.44	46.94	49.02	45.83	49.52	43.83	39.56	45.41	43.74
	+B2G	**57.71** +1.71	**26.40** +13.55	**51.45** +3.01	**48.85** +1.91	**51.48** +2.46	**49.50** +3.67	**53.19** +3.67	**47.51** +3.68	**46.79** +7.23	**48.72** +3.31	**50.16** +6.42

Table 3: Ablation Study Results. (a) Using the baseline feature extraction methods–Global-Local (GL) [45] and HybridGL (HGL) [18], all the subcomponents are integral for building B2G. **(b)** B2G consistently achieves higher performance for a number of top candidate masks (k: # of automatically generated clusters that differ per data sample). **(c)** Omitting negation or clustering reduces both mIoU and oIoU performance.

	CT	TC	SG	ViT-B/32 mIoU	ViT-B/32 oIoU	ViT-B/16 mIoU	ViT-B/16 oIoU
HGL				41.54	29.89	46.59	37.04
B2G	✓			**42.58**	**30.77**	**47.01**	**37.36**
GL				47.07	35.73	50.30	**38.35**
B2G	✓			**49.65**	**37.18**	**50.65**	38.24
GL		✓		47.70	36.32	50.68	**38.84**
B2G	✓	✓		**50.16**	**37.63**	**51.16**	**38.84**
GL	✓	✓		50.42	38.17	50.91	39.58
B2G	✓	✓	✓	**52.24**	**41.43**	**54.42**	**44.14**

(a) Subcomponent ablation

TC Mask #	GL+CT+SG mIoU	GL+CT+SG oIoU	B2G mIoU	B2G oIoU
2	50.47	40.28	**52.08**	**40.72**
3	50.79	41.29	**52.24**	**41.43**
4	49.42	40.81	**51.92**	**41.40**
k	–	–	**51.80**	**41.43**

(b) Comparison of zero-shot RIS performances using top candidates selected by the original and our candidate mask scores on RefCOCOg val (U) dataset

Neg.	Clus.	Refg	Ref	Ref+
		mIoU		
	✓	49.87	40.08	41.24
✓		50.36	41.40	41.50
✓	✓	**54.92**	**49.48**	**46.19**
		oIoU		
	✓	41.53	35.22	35.22
✓		41.16	35.32	34.53
✓	✓	**45.21**	**41.43**	**38.90**

(c) Effect of neg. and clus.

These work complementarily: `CT` refines language-aligned representations, while `TC` exploits spatial cues extracted from mid-layer features, together enabling more accurate spatial grounding.

`P-Map` design effects. We further examine the robustness of `P-Map` through top candidate mask quality (Tab. 3b). Across varying numbers of selected candidates (k), our method consistently outperforms Global-Local [45] (+`CT`+`SG`), demonstrating that `P-Map` produces reliable region proposals, with the top three masks yielding the best performance. Moreover, both negation and clustering are essential (Tab. 3c), as removing either step leads to a 4–11% drop, underscoring their importance for robust spatial grounding.

Multilingual module ablation. Text-side evaluations confirm that the language-induced geometric drift is robustly resolved by our multilingual mid-layer centroid injection, avoiding the hyperparameter sensitivity often observed in final-layer fusions. Detailed text-side ablations are provided in the Appendix. Additional translation-system and donor-language ablations in the Appendix show the same trend under both NLLB and GPT-based translations, and indicate that the gain is not explained solely by English/target-language hitchhiking or by the number of donor languages.

4 Related Work

Zero-shot RIS. While traditional RIS relies on costly pixel-level annotations [40, 44] or weakly-supervised training [31, 46], zero-shot methods circumvent task-specific training by leveraging pretrained VLEs in a proposal-and-rank pipeline: an off-the-shelf segmentor generates candidate masks [5, 10, 16, 38], and a frozen VLE ranks them via image–text similarity. Training-free improvements mainly focus on mask scoring through feature fusion (Cropping, Global-Local [45], HybridGL [18]), heuristic spatial reasoning (ReCLIP [32]), token-level similarity (Region Token [13]), or text augmentation (TAS [34]). Spatial guidance using diffusion models (Ref-Diff [22]) or Grad-CAM (CaR [33]; IteRPrimE [39]) requires additional models or gradient backpropagation. Crucially, these prior methods primarily rely on final-layer multimodal embeddings, which inherently suffer from spatial insensitivity and language-dependent shifts. In contrast, our framework bypasses this alignment bottleneck by extracting a novel spatial map (`P-Map`) directly from position-sensitive intermediate layers in a single forward pass. Recent dense VLE methods such as ClearCLIP [12] and ResCLIP [43] also exploit non-final CLIP representations for dense prediction. Our novelty is not the isolated observation that mid-layers can preserve localization cues, but a model- and modality-agnostic *probe* → *select* → *ground* framework that uses selected mid-layer cues for both spatial reranking and multilingual text stabilization at inference time.

Multilingual RIS. Extending RIS beyond English introduces additional challenges, as language-specific phrasing can distort cross-modal alignment. Multilingual benchmarks such as Room-Across-Room [11], xGQA [24], and COMFORT [48], along with large-scale resources like WIT [30] and XM3600 [35], enable multilingual evaluation and pretraining. Notably, many of these evaluation datasets are constructed by translating English-centric benchmarks, which inherently introduces translation-induced semantic drift [23]. Prior works typically mitigate such biases through resource-intensive pretraining and distillation (MURAL [8], UC2 [49], mCLIP [3]) or by developing new multilingual backbones (SigLIP [47], SigLIP2 [37]). In contrast, our method improves multilingual grounding at test time by probing frozen VLEs to identify language-stable intermediate layers and using a multilingual centroid to stabilize cross-modal similarity.

5 Conclusion

We revisit zero-shot referring image segmentation through the lens of layer-wise representational bias in pretrained Vision–Language Encoders. Our analysis shows that final-layer multimodal embeddings, though optimized for global semantic alignment, suppress structurally informative signals. This results in weakened positional sensitivity in vision embeddings and language-dependent geometric shifts in multilingual text embeddings. By probing intermediate layers and selectively leveraging structurally preserved representations, we mitigate both spatial and cross-lingual gaps without architectural changes or fine-tuning. Beyond RIS, the identified mid-layer pathway may also benefit other multimodal tasks that require fine-grained structural reasoning. A deeper theoretical understanding of how alignment objectives shape layer-wise geometry could further enable bias-aware use of pretrained VLEs.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (No. RS-2025-00520207); Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (Nos. RS-2024-00457882, 2022-0-01045, 2022-0-00680, RS-2019-II190075); the Advanced GPU Utilization Support Program funded by the Government of the Republic of Korea (Ministry of Science and ICT); the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No. RQT-25-120217); a grant partly supported by both IITP (MSIT) and Korea Evaluation Institute of Industrial Technology (KEIT) (MOTIE) (No. RS-2025-02217259); and the next-generation CultureTech technology development (R&D) project for cultural space AX transformation through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2026 (No. RS-2026-25508598).

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments (2018), https://arxiv.org/abs/1711.07280
2. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3828–3837 (2024)
3. Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., Wang, W.: mCLIP: Multilingual CLIP via cross-lingual transfer. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13028–13043. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.728, https://aclanthology.org/2023.acl-long.728/
4. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments (2020), https://arxiv.org/abs/1811.12354
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
7. Han, Z., Zhu, F., Lao, Q., Jiang, H.: Zero-shot referring expression comprehension via structural similarity between images and captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14364–14374 (2024)
8. Jain, A., Guo, M., Srinivasan, K., Chen, T., Kudugunta, S., Jia, C., Yang, Y., Baldridge, J.: Mural: Multimodal, multitask retrieval across languages (2021), https://arxiv.org/abs/2109.05125
9. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? investigating their struggle with spatial reasoning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding (2020), https://arxiv.org/abs/2010.07954
12. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: European Conference on Computer Vision (2024)
13. Li, J., Shakhnarovich, G., Yeh, R.A.: Adapting clip for phrase localization without further training. arXiv preprint arXiv:2204.03647 (2022)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
15. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv e-prints pp. arXiv–2304 (2023)

16. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
18. Liu, T., Li, S.: Hybrid global-local representation with augmented spatial guidance for zero-shot referring image segmentation. arXiv preprint arXiv:2504.00356 (2025)
19. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7086–7096 (2022)
20. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
21. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 792–807. Springer (2016)
22. Ni, M., Zhang, Y., Feng, K., Li, X., Guo, Y., Zuo, W.: Ref-diff: Zero-shot referring image segmentation with generative models. arXiv preprint arXiv:2308.16777 (2023)
23. Nogueira, F., Bernardino, A., Martins, B.: Comprehension of multilingual expressions referring to target objects in visual inputs (2025), https://arxiv.org/abs/2511.11427
24. Pfeiffer, J., Geigle, G., Kamath, A., Steitz, J.M.O., Roth, S., Vulić, I., Gurevych, I.: xgqa: Cross-lingual visual question answering (2022), https://arxiv.org/abs/2109.06082
25. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics (2020)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450 (2016)
28. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation (2021), https://arxiv.org/abs/2109.12098
29. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks (2020), https://arxiv.org/abs/1912.01734
30. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2443–2449. SIGIR '21, ACM (Jul 2021). https://doi.org/10.1145/3404835.3463257, http://dx.doi.org/10.1145/3404835.3463257
31. Strudel, R., Laptev, I., Schmid, C.: Weakly-supervised segmentation of referring expressions. arXiv preprint arXiv:2205.04725 (2022)

32. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. arXiv preprint arXiv:2204.05991 (2022)
33. Sun, S., Li, R., Torr, P., Gu, X., Li, S.: Clip as rnn: Segment countless visual concepts without training endeavor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13171–13182 (2024)
34. Suo, Y., Zhu, L., Yang, Y.: Text augmented spatial aware zero-shot referring image segmentation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 1032–1043. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.73, https://aclanthology.org/2023.findings-emnlp.73/
35. Thapliyal, A.V., Pont-Tuset, J., Chen, X., Soricut, R.: Crossmodal-3600: A massively multilingual multimodal evaluation dataset (2022), https://arxiv.org/abs/2205.12522
36. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
37. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
38. Wang, X., Yu, Z., De Mello, S., Kautz, J., Anandkumar, A., Shen, C., Alvarez, J.M.: Freesolo: Learning to segment objects without annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14176–14186 (2022)
39. Wang, Y., Ni, J., Liu, Y., Yuan, C., Tang, Y.: Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8159–8168 (2025)
40. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
41. Wu, C., Lin, Z., Cohen, S., Bui, T., Maji, S.: Phrasecut: Language-based image segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10216–10225 (2020)
42. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
43. Yang, Y., Deng, J., Li, W., Duan, L.: Resclip: Residual attention for training-free dense vision-language inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
44. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18155–18165 (2022)
45. Yu, S., Seo, P.H., Son, J.: Zero-shot referring image segmentation with global-local context features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19456–19465 (2023)
46. Yu, S., Seo, P.H., Son, J.: Pseudo-ris: Distinctive pseudo-supervision generation for referring image segmentation. In: European Conference on Computer Vision. pp. 18–36. Springer (2024)

47. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
48. Zhang, B., Fei, Z., Xiao, Z., Wang, Y., Dong, J.: Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In: The Twelfth International Conference on Learning Representations (2024)
49. Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z., Liu, J.: Uc2: Universal cross-lingual cross-modal vision-and-language pre-training (2021), https://arxiv.org/abs/2104.00332