arXiv:2608.30621v1 [cs.CV] 31 Aug 2026

Cost-efficient Active Learning for Referring Image Segmentation and Grounding

Junbeom Hong[*,1] **Seonghoon Yu**[*,2] **Hyung Rok Jung**[3]
Sundong Kim[3] **Jeany Son**[4]

[1]Meissa [2]KAIST [3]GIST [4]POSTECH

jbhong@meissa.ai seonghoon.yu@kaist.ac.kr jeany@postech.ac.kr

Abstract

Collecting natural-language referring expressions along with region annotations, such as masks or boxes, is a major bottleneck in visual grounding (VG), as annotators must write descriptions that distinguish target regions from visually similar ones. We tackle this by formulating active learning (AL) for VG under the realistic setting where only raw images are available without accompanying text. Since ground-truth text is unavailable, sample selection must estimate which images contain ambiguous regions that would require discriminative referring expressions. To address this, we generate auxiliary region-text pairs using foundation models, and introduce Referred Region Ambiguity, a new acquisition function that measures whether the model's confidence collapses onto a single region or disperses across multiple candidates. It allows our method to prioritize images with strong cross-region competition, which are more informative due to their visual ambiguity. We also design a referring-expression annotation interface that helps annotators quickly focus on writing discriminative language with a few clicks. Experiments on RIS and REC benchmarks show that our AL framework consistently outperforms several AL baselines, while a user study shows up to 1.6× faster description labeling of ours. Code is available at https://github.com/junbum766/ALRIS.

1 Introduction

Deep learning models often require large amounts of high-quality human annotations, making data collection a major bottleneck. Active learning (AL) addresses this issue by selecting the most *informative* unlabeled samples for annotation, reducing labeling costs while maintaining performance. It has been widely applied to computer vision tasks such as image classification (Gal et al., 2017; Sener and Savarese, 2018), object detection (Yang et al., 2024; Lyu et al., 2023), and semantic segmentation (Cai et al., 2021; Hwang et al., 2023).

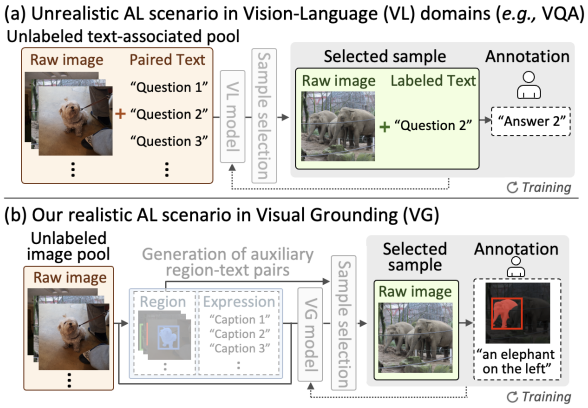


Figure 1: Illustration of AL settings in vision-language domains: (a) existing *unrealistic* scenario (Kim et al., 2021; Lin and Parikh, 2017) in VQA assumes that costly textual questions related to each image are already available in the unlabeled pool, whereas (b) our *realistic* setting in VG considers a practical scenario where only raw images are provided without textual descriptions.

Despite its success in vision-only tasks, active learning (AL) remains underexplored for text-conditioned vision-language (VL) tasks such as visual question answering (VQA) and visual grounding (VG). Existing AL studies mainly focus on VQA and assume that image-text pairs are already available in the unlabeled pool (Karamcheti et al., 2021; Lin and Parikh, 2017; Kim et al., 2021), requiring annotators only to provide answer labels (Fig. 1a). This assumption is often unrealistic: real-world unlabeled data typically consists of raw images, where both linguistic labels, such as questions in VQA or referring expressions in VG, and their corresponding labels (answers or referred regions) must be annotated. Moreover, AL for VG tasks such as referring image segmentation (RIS) and referring expression comprehension (REC) has not been explored despite their high annotation cost.

Another challenge in VG is that informativeness is characterized differently than in standard vision-

[*]Equal contribution.



(a) Informative example (b) Non-informative example

Figure 2: Illustration of informativeness in VG: (a) the model assigns high confidence to multiple regions, indicating an informative sample, while (b) a referred one is solely activated by the model, showing low informativeness.

only tasks. In VG, the model must pinpoint one referred region among many visually or semantically similar candidates. Therefore, when the model assigns high confidence to multiple regions simultaneously, such samples are, in fact, highly informative: they reveal exactly where the model lacks fine-grained discriminative ability (Fig. 2). However, existing acquisition functions designed for segmentation or object detection fail to capture this property, since they primarily select samples with moderate overall confidence (*e.g.*, pixel entropy (Mackowiak et al., 2018), box-level confidence (Choi et al., 2021)). As a result, cases where the model strongly activates multiple regions are mistakenly treated as low-informative, although they correspond to informative VG samples.

In this paper, we introduce a new active learning task for visual grounding that selects and annotates samples directly from raw images, reducing manual effort by labeling only a small number of highly informative samples (Fig. 1b). First, to address the lack of textual descriptions in the unlabeled image pool, we generate auxiliary region-text pairs, providing the required multi-modal input for a VG model and sample selection. Second, to capture VG-specific informativeness, we introduce a novel acquisition function termed *Referred Region Ambiguity*. This metric measures how the model's confidence is distributed across multiple candidate regions instead of being concentrated in one referred region. Its high ambiguity indicates that multiple regions are similarly plausible under the same expression, exposing genuine ambiguity that is worth labeling. By prioritizing such ambiguous samples, our framework enables the VG model to rapidly improve its fine-grained discriminative capability with minimal annotation costs.

Furthermore, we propose an efficient referring-expression annotation interface along with mask annotations for VG. Our interface first lets annotators specify the target region, then assists expression writing with region-aware top-k word suggestions. Rather than typing full sentences from scratch, annotators construct distinct referring expressions by sequentially selecting suggested top-k words relevant to the target region with a few clicks, achieving a significant reduction in the labeling cost.

Under the realistic AL scenarios in VG, our active learning framework achieves state-of-the-art performance over several AL baselines on both RIS and REC benchmarks. Notably, the proposed labeling tool achieves up to 1.6× faster labeling speeds for referring expression annotations, compared to the naïve labeling in a user study.

Our contributions can be summarized as follows:

- We introduce a new task setting for active learning in visual grounding, where sample selection must be performed from raw images without pre-existing image-text pairs, reflecting a realistic data collection scenario.
- We propose a novel acquisition, *Referred Region Ambiguity*, built on auxiliary region-text pairs, which selects ambiguous samples where the model assigns high confidence to multiple regions rather than a referred one, enabling faster learning of fine-grained discrimination.
- We present an efficient referring-expression labeling tool with region annotations that reduces annotation cost by reformulating expression construction as a word-by-word click process.
- Our approach surpasses several AL baselines on RIS and REC benchmarks while concurrently delivering superior labeling efficiency in both speed and annotation cost.

2 Related Work

Active Learning. Active learning (AL) reduces annotation costs by selecting informative samples from unlabeled data pools. Prior work has explored uncertainty (Wang and Shang, 2014; Gal et al., 2017; Yoo and Kweon, 2019; Gwon et al., 2025), diversity (Tran et al., 2019), and hybrid (Zhdanov, 2019; Shui et al., 2020) criteria in image classification, and then extended to detection (Choi et al., 2021; Lyu et al., 2023), segmenta-

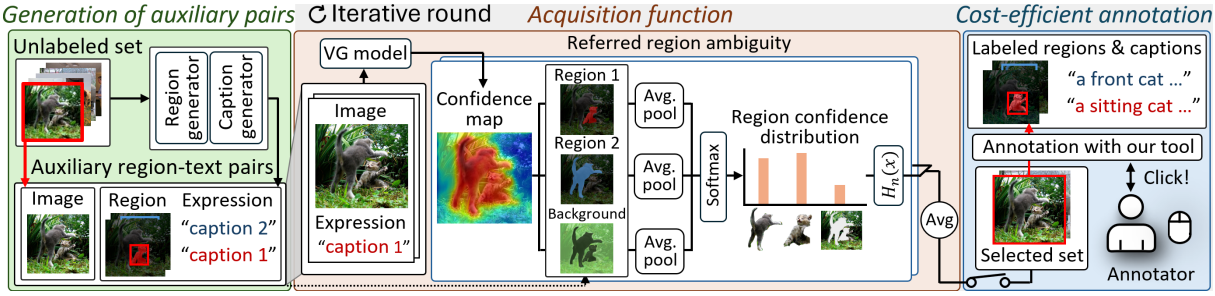


Figure 3: The proposed AL framework for visual grounding. We first generate auxiliary region-text pairs for all raw images in the entire unlabeled pool (Sec. 3.2). Following this, iterative rounds of sample selection and labeling begin. In each round, our referred region ambiguity (Sec. 3.3) scores the informativeness of all raw images by utilizing their auxiliary region-expression pairs. The top-scoring images are then sequentially annotated via our cost-efficient interface (Sec. 4) until the round budget is exhausted.

tion (Popp et al., 2025; Kim et al., 2023b, 2024a), and VQA (Karamcheti et al., 2021; Lin and Parikh, 2017; Kim et al., 2021). However, VL AL typically assumes pre-existing language inputs. Although ACTRESS (Kang et al., 2024) actively selects pseudo labels for semi-supervised VG from image-text pairs, it does not address human annotation cost incurred in text modality. In contrast, we study realistic active VG data collection from raw images, where both target regions and referring expressions are newly annotated.

Cost-efficient Annotation Tool. Reducing per-sample labeling cost is crucial for AL efficiency. Prior works have reduced visual annotation costs for segmentation using superpixels (Colling et al., 2020; Hwang et al., 2023) or foundation models (Kim et al., 2024a; Popp et al., 2025), but mainly target mask annotation. In contrast, referring expression annotation remains largely unexplored despite a bottleneck in VG. To this end, we introduce a first interface for referring expressions, together with referred region annotations.

Visual Grounding. Visual grounding (VG) localizes regions from language, producing masks for RIS (Kim et al., 2024b; Yu et al., 2024a) or boxes for REC (Deng et al., 2023; Liu et al., 2023). Since VG annotation requires both regions and detailed expressions, prior work has reduced supervision via weakly-(Kim et al., 2023a; Jin et al., 2023), semi-supervised (Sun et al., 2023; Zang et al., 2024), zero-shot (Yu et al., 2023), and pseudo-labeling (Jiang et al., 2022; Yu et al., 2024b) approaches. However, their gap to fully supervised methods (Yu et al., 2025; Xiao et al., 2024) highlights the value of human labels. In this work, we explore active learning for VG to minimize annotation cost while preserving human-label quality.

3 Active Learning for Visual Grounding

We introduce an active learning framework for visual grounding that reduces manual labeling costs by selecting a small subset of informative raw images. Our goal is to identify challenging samples where the model struggles to ground an expression to the intended region among similar candidates. Such ambiguity reveals the fine-grained discrimination required for VG. We describe the task setup, auxiliary region-text pair construction, and our acquisition function in this section.

3.1 Overall Framework

We start by formalizing the active learning task in the context of visual grounding under a realistic setting, where only raw images are available in an unlabeled pool. Formally, let $\mathcal{I} = \{I_1, \ldots, I_{|\mathcal{I}|}\}$ denote an unlabeled image set. The goal is to iteratively select a subset of informative images $\{I_n\}_{n=1}^{N_r} \subset \mathcal{I}$ for human labeling at each round r. To enable VG sample selection on raw images, our framework begins with *generation of auxiliary region-text pairs* (Sec. 3.2), where it produces auxiliary pairs for each text-less image to infer the informativeness.

In every round r, our *acquisition function* (Sec. 3.3) assesses informativeness scores of every unlabeled image using its corresponding auxiliary samples. Based on these scores, the top informative raw images $\{I_n\}_{n=1}^{N_r} \subset \mathcal{I}$ are sequentially annotated by humans via our *cost-efficient interface* (Sec. 4) until the round's budget B_r is exhausted. These newly labeled samples $\mathcal{D}_r$ are added to the manually labeled set $\bigcup_{i=1}^{r-1} \mathcal{D}_i$, and then the VG model θ_{r-1} is retrained on these updated labeled data $\bigcup_{i=1}^{r} \mathcal{D}_i$ to begin the next AL round. The overall procedure is summarized in Algorithm 1 and illustrated in Fig. 3. In the following sections, we further elaborate on each step in more detail.

Algorithm 1 Proposed Active Learning Framework for VG

Require: Unlabeled image set $\mathcal{I}$, the number of AL rounds R, budget per round B_r, an initial model θ_0

Build a set of auxiliary pairs $\mathcal{D}_{\text{aux}}$ on $\mathcal{I}$ ▷ Sec. 3.2
for $r = 1, 2, \ldots, R$ **do**
  Score informativeness of all raw images $\forall I \in \mathcal{I}$ via θ_{r-1} and $\mathcal{D}_{\text{aux}}$ ▷ Sec. 3.3
  Annotate top-scored images with our annotation interface ▷ Sec. 4
  Build a set of labeled images $\mathcal{D}_r$ within budget B_r
  Get the model θ_r trained on labeled dataset $\bigcup_{i=1}^{r} \mathcal{D}_i$
end for
Return The final model θ_R

3.2 Generation of Auxiliary Region-Text Pairs

A core challenge in applying AL to VG lies in estimating raw-image informativeness without textual descriptions. We address this by generating a set of auxiliary region-text pairs for each unlabeled image, offering potential semantics to estimate its informativeness. Specifically, for a given image I, we extract a set of N potential regions $\{r_i^{\text{aux}}\}_{i=1}^{N}$ using Grounded-SAM (Ren et al., 2024) guided by COCO 80 categories as semantic prompts, where N depends on the number of detected regions.

Next, we generate corresponding auxiliary captions $\{t_i^{\text{aux}}\}_{i=1}^{N}$ using ViP-LLaVA (Cai et al., 2024) for all regions, with each region highlighted by a red box to guide generation. These region-caption pairs form $\mathcal{D}_{\text{aux}}$ over the unlabeled pool $\mathcal{I}$, which provide potential signals to score the informativeness of each raw image via the acquisition function in the next subsection.

3.3 Acquisition Function

Our acquisition function, termed **Referred Region Ambiguity** (Fig. 3), is designed to capture the unique source of informativeness in VG: ambiguity among candidate referred regions. While standard acquisition functions measure uncertainty at the pixel or box level (Mackowiak et al., 2018; Choi et al., 2021), they overlook the cross-region competition induced by a referring expression. In contrast, our method treats candidate regions as competing referents and computes a region-level confidence distribution for each auxiliary caption. It assigns high scores when the model distributes confidence across multiple plausible regions, revealing its difficulty in identifying the intended referent, and low scores when confidence is concentrated on a single region. This formulation directly targets the fine-grained discrimination required in VG and provides a more faithful informativeness signal than generic vision-only acquisition criteria that do not account for language-conditioned region ambiguity.

Given a raw image I with its corresponding N auxiliary pairs $\{(r_i^{\text{aux}}, t_i^{\text{aux}})\}_{i=1}^{N}$, where r_i^{aux} is an auxiliary region (*i.e.*, a mask for RIS or a box for REC) and t_i^{aux} is its corresponding caption, we quantify how exclusively the model attends to the intended region for each auxiliary caption. For each caption t_i^{aux}, the model produces a confidence map (*i.e.*, a final logit scoring map before thresholding for RIS or attention map from the final decoder layer for REC), $p^i = \theta(I, t_i^{\text{aux}}) \in \mathbb{R}^{h \times w}$. We then compute region-level confidence scores s_j^i over $N+1$ regions (N auxiliary regions plus a background region r_0^{aux}) by averaging the confidence values within each region, as follows:

$$s_j^i = \frac{1}{|r_j^{\text{aux}}|} \sum_{(x,y) \in r_j^{\text{aux}}} p_{(x,y)}^i \in \mathbb{R}, \ \forall j \in \{0, \ldots, N\},$$

where $p_{(x,y)}^i$ denotes the confidence score at the pixel coordinate (x, y). These scores are normalized with a softmax over all regions, producing a probability distribution $\mathbf{q}^i = \frac{\exp(s_j^i)}{\sum_{j=0}^{N} \exp(s_j^i)} \in \mathbb{R}^{N+1}$ for the i-th caption. We then compute the normalized entropy $H_n(\mathbf{q}^i)$ of $\mathbf{q}^i$, as follows:

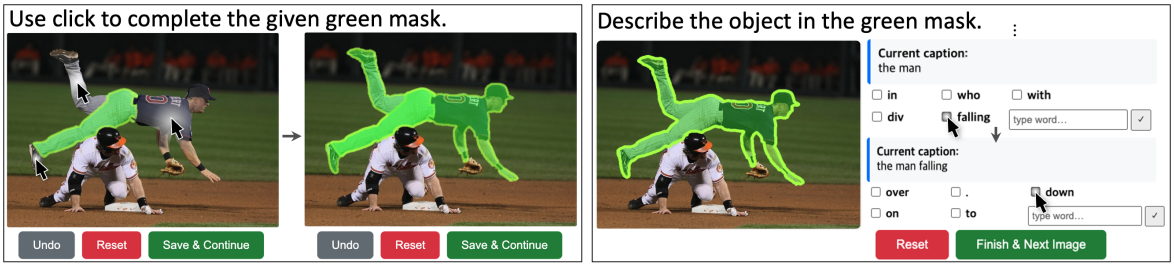
$$H_n(\mathbf{q}^i) = -\frac{\sum_{j=0}^{N} q_j^i \log(q_j^i)}{\log(N+1)} \in \mathbb{R}, \qquad (1)$$

where $H_n(\mathbf{q}^i)$ is computed by taking the entropy of $\mathbf{q}^i$ and then dividing it by the maximum entropy value over $N+1$ regions, $\log(N+1)$. A higher value of $H_n(\mathbf{q}^i)$ indicates greater ambiguity, meaning the model assigns similar confidence to multiple candidate regions rather than focusing on a single dominant one.

Finally, the informativeness of an image I is then computed by averaging $H(\mathbf{q}^i)$ over all its auxiliary captions, as follows:

$$E(I) = \frac{1}{N} \sum_{i=1}^{N} H_n(\mathbf{q}^i), \qquad (2)$$

which we define as *Referred Region Ambiguity*.



(a) Target-Mask Specification (b) Referring-Expression Labeling Interface

Figure 4: The proposed cost-efficient annotation interface. (a) target-mask specification, where the user refines a presented mask by clicking on a desired region in sequence, and (b) text labeling interface, where the annotator composes a referring expression for the green mask by clicking on suggested word candidates sequentially, with the option to manually type a word if none of the suggested words are suitable. Demo videos are available in our GitHub repository.

Images with the highest $E(I)$ scores are then annotated with our cost-efficient labeling tool (Sec. 4) until the next AL round budget is exhausted.

4 Cost-efficient Annotation Interface

To mitigate annotation overhead in our AL pipeline, we further present a cost-efficient tool for labeling referring expressions. As shown in Fig. 4, it enables click-driven target-region annotation and word-by-word expression construction via simple clicks, avoiding labor-intensive mask drawing and free-form text typing. We first detail our annotation workflow and then present its cost measurement.

Target-Region Specification. We first specify the target region to be described by a referring expression. For RIS, annotators start from an auxiliary mask from Sec. 3.2 and refine it through simple clicks, each selecting and merging a pre-extracted SAM (Kirillov et al., 2023) mask. For REC, we use two-click bounding-box specification.

Referring-Expression Annotation. For each finalized region, annotators construct a referring expression word by word, as shown in Fig. 4b. At each step, the interface presents k suggested words, which annotators can select with a click or replace by choosing the *typing* option. This process repeats until the expression is complete. The word candidates are generated by a region multimodal LLM (*i.e.*, ViP-LLaVA (Cai et al., 2024)): at each auto-regressive decoding step, the model predicts the next-word distribution conditioned on the image, region, and previously selected words, and the top-k words are shown as clickable candidates.

Cost Measurement. We define each sample's labeling cost as the sum of its region cost (mask for RIS or box for REC) and text cost, computed as:
1) **Target-region costs** (C_{region}): for RIS, it is defined as the number of clicks required to finalize the mask labeling; for REC, it is fixed to two clicks to specify the diagonal corners of a single box.
2) **Referring-expression costs** (C_{exp}): we define this cost as the sum of each word selection cost C_{w} over an expression. For each word selection, if the desired word appears in the top-k suggestions with probability p_k, the user selects it from $k+1$ options, costing $\log_2(k+1)$ bits. Otherwise, the user selects the typing option with the same cost and additionally pays a typing cost C_{type}. C_{type} is computed as the conditional information content of typed letters, estimated by a frequency-based letter-level 3-gram model. Formally,

$$C_{\text{word}} = \underbrace{\log_2(k+1)}_{\text{Click a word or typing option}} + \underbrace{(1-p_k)C_{\text{type}}}_{\text{Type a word manually}},$$

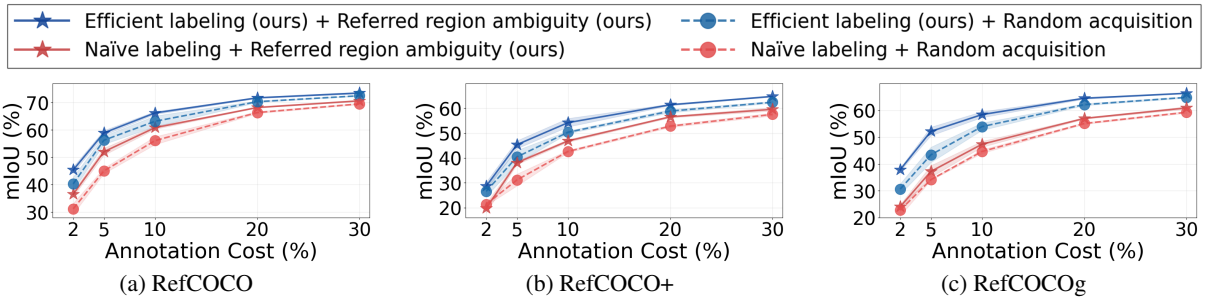
$$\text{where } C_{\text{type}} = -\sum_{i=1}^{n} \log_2 p_l(l_i|l_{i-1}, l_{i-2}), \quad (3)$$

and $p_l(l_i|l_{i-1}, l_{i-2})$ is the conditional probability of letter l_i given previous two letters. Statistics for p_k and $p_l(l_i|l_{i-1}, l_{i-2})$ are in Appendix B.1.

5 Experiments

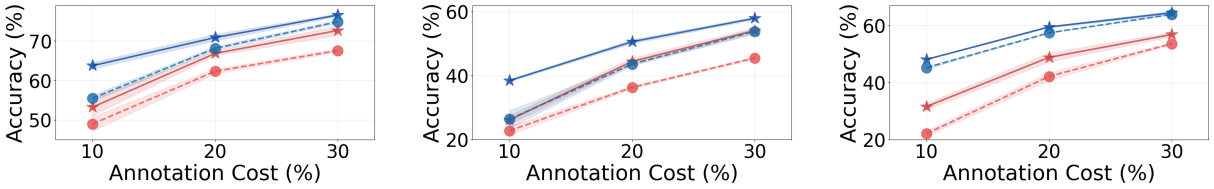
5.1 Implementation Details

For generation of auxiliary region-text pairs (Sec. 3.2), We extract potential regions from Grounded-SAM (Ren et al., 2024) using a confidence threshold of 0.23, and limit the number of extracted ones per image to 7. We then prompt ViP-LLaVA (Cai et al., 2024) with the instruction "*describe the object in the red box*" to generate the corresponding expression. **For acquisition function** (Sec. 3.3), the background region is obtained as the remaining area after excluding all extracted instance regions. **For cost-efficient annotation interface** (Sec. 4), we set the number of suggested word candidates to $k = 5$. The annotation interface runs on a single RTX 3090 GPU. **For the experiments**, we train DETRIS-B (Huang et al., 2025) model for RIS and MaPPER-B model (Liu



(a) RefCOCO (b) RefCOCO+ (c) RefCOCOg

Figure 5: Comparison with the naïve labeling and random acquisition on a **RIS** task.



(a) RefCOCO (b) RefCOCO+ (c) RefCOCOg

Figure 6: Comparison with the naïve labeling and random acquisition on a **REC** task. The reported results are averaged on the val and test (A,B) splits of each dataset. The annotation cost (x-axis) is calculated as the measured cost of the acquired samples via used labeling tool divided by the total manual annotation cost of the dataset.

et al., 2024) for REC on two A6000 GPUs. In both models, we follow the training configurations as detailed in their respective models. Following previous AL works (Munjal et al., 2022), we train the initial model on 1% random samples to address the cold-start problem (Mahmood et al., 2021). The GPU costs incurred by the auxiliary generation and the annotation interface are in Appendix A.5.

5.2 Dataset and Metric

For dataset, we evaluate our active learning framework on standard benchmarks for visual grounding tasks (RIS and REC): RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), RefCOCOg (Mao et al., 2016). Details on each dataset are in Appendix. For evaluation metric, we use standard metrics for both RIS and REC. For RIS, we adopt mIoU (mean IoU), which is the average of the per-sample IoU scores. For REC, we employ Prec@50, which is an accuracy only when IoU is over 0.5.

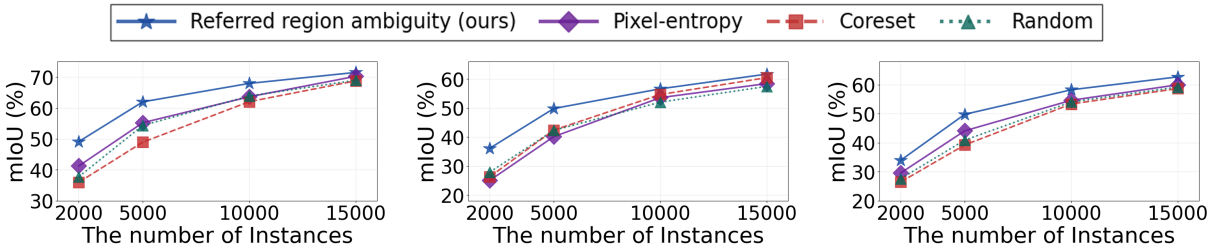
5.3 Main Results

Referring Image Segmentation. In Fig. 5, we compare the performance of our active learning framework with random acquisition and naïve labeling (*i.e.*, polygon drawing for masks and manual typing for expressions) on RIS benchmark under annotation budgets of 2%, 5%, 10%, 20%, and 30%. The x-axis denotes the measured annotation cost of acquired samples divided by the total naïve labeling cost of the dataset. For clarity, the reported mIoU results are averaged across the val and test(A/B) of each RefCOCO(+/g) and three independent runs to ensure statistical significance. The proposed acquisition outperforms random acquisition under both labeling settings (*i.e.*, our efficient tool and naïve tool), while our annotation interface delivers strong performance across all budgets, proving its efficiency.

Referring Expression Comprehension. We also present the performance of our method on REC compared to random acquisition and naïve labeling (*i.e.*, solely manual typing for expressions) under 10%, 20%, and 30% annotation budgets in Fig. 6. Results are averaged over the val and test(A/B) splits of RefCOCO(+/g) and three separate runs. Our acquisition consistently outperforms random selection, while our text annotation tool achieves higher performance than the naïve tool at the same annotation cost, demonstrating its cost efficiency.

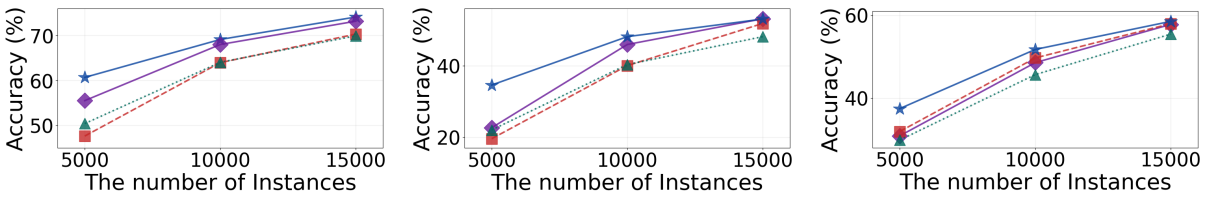
Comparisons on Acquisition Function. To isolate the effect of our acquisition, we compare it with other baselines on RIS and REC using the number of selected region-text pairs as the budget, removing the influence of labeling cost efficiency. We consider random selection, pixel-entropy (Mackowiak et al., 2018), and core-set (Sener and Savarese, 2018), detailed in Appendix B.2. As shown in Figs. 7 and 8, our referred region ambiguity consistently outperforms all baselines, demonstrating its effectiveness in selecting informative VG samples.

Verification of Annotation Tool with User Study. In Tab. 2, we empirically validate the efficiency of our proposed labeling tool via a user study with 30 annotators on Amazon Mechanical Turk. We



(a) RefCOCO (b) RefCOCO+ (c) RefCOCOg

Figure 7: Comparisons with several acquisition baselines on a **RIS** task.



(a) RefCOCO (b) RefCOCO+ (c) RefCOCOg

Figure 8: Comparisons with several acquisition baselines on a **REC** task. By comparing performance based on the number of acquired instance-text pairs, we isolate the effectiveness of each acquisition from the annotation tool.

Acquisition Function	Labeling Interface		Annotation Cost					Average
	Mask	Text	2%	5%	10%	20%	30%	
			$31.20_{\pm 1.98}$	$45.04_{\pm 0.97}$	$56.10_{\pm 1.44}$	$66.37_{\pm 0.34}$	$69.53_{\pm 0.25}$	53.65
✓			$36.53_{\pm 0.90}$	$52.05_{\pm 1.28}$	$60.84_{\pm 0.56}$	$68.20_{\pm 0.14}$	$70.65_{\pm 0.22}$	57.66
✓	✓		$38.08_{\pm 0.52}$	$54.09_{\pm 1.30}$	$62.71_{\pm 0.49}$	$69.83_{\pm 0.17}$	$71.92_{\pm 0.40}$	59.33
✓		✓	$38.66_{\pm 1.18}$	$56.40_{\pm 1.05}$	$63.42_{\pm 0.44}$	$70.01_{\pm 0.24}$	$72.37_{\pm 0.31}$	60.17
✓	✓	✓	$\mathbf{45.41}_{\pm 1.07}$	$\mathbf{58.94}_{\pm 0.86}$	$\mathbf{66.22}_{\pm 0.51}$	$\mathbf{71.74}_{\pm 0.19}$	$\mathbf{73.50}_{\pm 0.29}$	**63.16**

Table 1: Ablation study within proposed components. Additional analyses are provided in Appendix A.

Tools	Mask Labeling		Referring-Expression Labeling		
	Time (s)	mIoU	Time (s)	Sen.-BERT Sim.	CLIP Score
Naïve	60.6	73.0	45.8	39.9	19.2
Ours	9.9	80.4	28.6	44.4	22.2

Table 2: User study comparisons with the naïve labeling. We involve 30 annotators from Amazon Mechanical Turk. Time (s) means the average time per sample.

compare it with the naïve labeling approach (*i.e.*, polygon clicks for masks and manual typing for text), using per-sample time and quality metrics, including mIoU for masks, Sentence-BERT (Reimers and Gurevych, 2019) similarity between expressions and ground-truth one, and CLIP score (Hessel et al., 2021) between cropped images around masks and expressions. Our interface achieves faster annotation while maintaining high-quality mask specification and referring-expression labeling. Examples of user-annotated samples for each tool are visualized in Fig. 16 of Appendix C.1.

5.4 Ablation Study

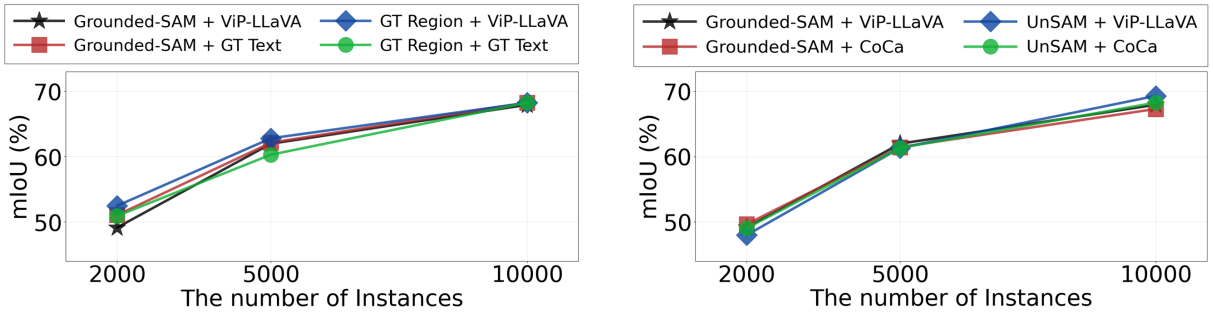
We conduct ablation studies on RefCOCO for RIS, reporting average mIoU over the val, test(A/B) splits to validate the effectiveness of our approach.

Effect of Each Proposed Component. Tab. 1 presents the contribution of each component in our AL framework. Starting from random acquisition with naïve annotation, replacing random selection with our referred region ambiguity notably improves performance, especially under low budgets. Adding efficient mask and text-labeling tools further reduces annotation costs and improves performance, with their combination yielding the best results. Additional analyses (*e.g.*, auxiliary regions with other class vocabularies and GPU cost) are in Appendix A.

Ablations	Annotation Budget			Measured Cost over Dataset
	2%	10%	20%	
(a) Mask Annotations				
Naïve polygon	38.66	63.42	70.01	1,616,155
Superpixel	43.94	64.80	71.08	527,770
Pre-extracted SAM (ours)	**45.41**	**66.22**	**71.74**	**53,970**
(b) Text Annotations				
Naïve typing	38.08	62.71	69.83	6,493,300
Sentence selection	40.18	64.34	70.16	**2,814,811**
Word selection (ours)	**45.41**	**66.22**	**71.74**	4,238,677

Table 3: Ablation on the annotation interface. Measured cost denotes the total labeling cost required to label all ground-truth data in RefCOCO with each tool.

Ablation in Annotation Interface. To confirm the effectiveness of the proposed labeling interface, Tab. 3 compares our tool with several baseline labeling strategies. **For mask labeling**, we compare 1) *naïve polygon*, where the user clicks a series of vertices around region boundaries, and 2) *super-*



(a) Ablation with GT annotations

(b) Ablation with other generators

Figure 9: Ablation on the auxiliary generator. (a) we use GT labels as auxiliary pairs and (b) we use UnSAM (Wang et al., 2024) (unsupervised model) for region generator and CoCa (Yu et al., 2022) (trained solely on web-crawled data without human annotations) for text generator.

Quality	Auxiliary Region Generator			Auxiliary Text Generator		
	GT mask	Grounded-SAM	UnSAM	GT text	ViP-LLaVA	CoCa
mIoU or Sim.	100	83.56	56.85	100	54.72	32.97

Table 4: Quality of auxiliary pairs generated from the auxiliary generators in the ablation (Fig. 9b). For the region quality, we use mIoU with the ground-truth masks. For the text quality, we use Sentence-BERT (Reimers and Gurevych, 2019) similarity with the ground-truth text.

pixel (Van den Bergh et al., 2012), where the user selects over-segmented superpixels to produce a mask. Compared to them, our mask tool, based on pre-extracted SAM masks, achieves the best results across all budgets and the lowest mask click cost, demonstrating its superiority. **For text labeling**, we compare 1) *manual typing*, where annotators manually write full sentences from scratch, and 2) *sentence selections*, where annotators select the most suitable expression among several MLLM-generated candidates. Our word-level selection approach with MLLM-guided decoding achieves the highest performance with the reduced manual effort over all compared methods, enabling more high-quality and faster text labeling, as validated in the user study (Tab. 2).

Ablation on Auxiliary Generator. To examine the sensitivity of our active learning framework to auxiliary generators, we conduct ablations using diverse auxiliary sources in Fig. 9. When using the ground-truth annotations as auxiliary region-text pairs (Fig. 9a), our AL method consistently remains robust. For the other generators (Fig. 9b), we use UnSAM (Wang et al., 2024) (unsupervised model) for region generator and CoCa (Yu et al., 2022) (with the weights trained solely on web-crawled data without human annotations) for text generation. Despite quality differences among auxiliary sources (Tab. 4), our active learning framework consistently maintains robust performance. These results indicate that our approach is not sensitive to the auxiliary generators. We attribute this robustness to the role of auxiliary pairs as candidate proposals for acquisition rather than final supervision; active learning mainly depends on the relative ranking of informative candidates, so moderate noise in masks or text often does not substantially change the priority of selection.

5.5 Qualitative Analysis

Analysis on Selected Samples via Acquisition. To highlight that our acquisition selects more complex VG images, we present qualitative examples of selected images in Fig. 10 and a statistical analysis in Fig. 11 compared with pixel-entropy (Mackowiak et al., 2018) and coreset (Sener and Savarese, 2018). Our method often selects images with visual similar regions, which is a hard scenario that can confuse the model, and thus promotes its discrimination capability. For a statistical comparison, we measure the feature similarity between the referred region and all other regions within an image, and then average this over all selected samples. The high similarity achieved by our acquisition confirms the presence of many visually similar distractors in the selected samples, indicating that ours effectively prioritizes challenging VG samples.

Analysis of Referred Region Ambiguity. In Fig. 12, we visualize the process of computing our informativeness score and compare it with the pixel-entropy (Mackowiak et al., 2018). Pixel-entropy estimates the informativeness by averaging the entropy values (visualized in the figure) across all pix-



(a) Ours (b) Pixel-entropy (c) Coreset

Figure 10: Visualization of selected images compared with other acquisition functions. Paired ground-truth region-text annotations are denoted by matching colors. More visualization are present in Fig. 20 of Appendix C.4.

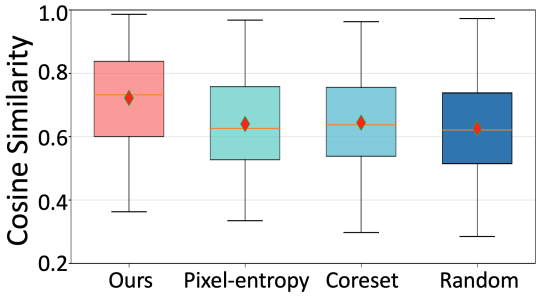


Figure 11: Statistical comparison of selected pool with other methods. The measured similarity is the average feature similarity between the referred region and all other regions within the image, calculated across all samples in the selected pool.



Figure 12: Qualitative analysis on the process of computing the informativeness score compared to pixel-entropy. The pixel-entropy displays a raw entropy map derived from the confidence map, while ours shows the measured region-wise entropy computed from region-wise confidence distribution.

els in the confidence map. In contrast, our referred region ambiguity quantifies the region-level ambiguity (also visualized in the figure) from region-wise confidence scores, and then averages these region-wise entropy values. By focusing on region-level scores, our approach effectively selects complex yet informative VG examples. More visualization are provided in Fig. 18 of Appendix C.5.

6 Conclusion

In this paper, we present a novel active learning framework for visual grounding that operates under the realistic setting of a purely raw-image pool by utilizing auxiliary region-text samples as potential signals. By estimating VG-specific informativeness via our tailored acquisition and simplifying the labor-intensive labeling process via the proposed interface, our approach significantly reduces overall manual effort while consistently outperforming all AL baselines on RIS and REC benchmarks.

7 Limitations

While our annotation tool reduces human labor costs, it introduces additional GPU overhead. In particular, generating word-level suggestions with multimodal LLMs during annotation requires extra GPU memory and computation, as analyzed in Tab. 9 of Appendix A.5. Since our annotation tool is designed to reduce human labeling costs, we do not explicitly consider the GPU costs incurred by the tool in its design. However, these costs are marginal, as the interface runs on a single RTX 3090 and requires only modest computation compared to the saved human labor. Overall, our tool reduces the annotation cost from $18,296 with naïve labeling to $6,752, achieving a 2.7× cost reduction. We leave GPU-aware annotation-tool design for future work.

8 Ethical Consideration

Our work aims to reduce human annotation burden in visual grounding. Since our annotation tool involves human annotators, they should be clearly informed of the task and fairly compensated. Although our interface reduces repetitive manual effort, human verification remains important to ensure accurate and unbiased referring expressions.

Acknowledgments

This work was supported by the IITP grants (RS-2019-II191906 (5%), 2019-0-01842 (5%), RS-2022-II220926 (35%), RS-2026-25518317 (30%)), the NRF grant (RS-2025-24535146 (10%)), the InnoCORE program (N10250156 (10%)), GIST (Future-leading Specialized Research Project (5%)), and the AI Computing Support Project for R&D, funded by MSIT, Korea.

References

Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. 2021. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. 2021. Active learning for deep object detection via probabilistic modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pascal Colling, Lutz Roese-Koerner, Hanno Gottschalk, and Matthias Rottmann. 2020. Metabox+: A new region based active learning method for semantic segmentation using priority maps. *arXiv preprint arXiv:2010.01884*.

Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. 2022. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *IEEE International Conference on Image Processing (ICIP)*.

Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. 2023. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(11):13636–13652.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research (TMLR)*.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*.

Yeho Gwon, Sehyun Hwang, Hoyoung Kim, Jungseul Ok, and Suha Kwak. 2025. Enhancing cost efficiency in active learning with candidate set query. *Transactions on Machine Learning Research (TMLR)*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. 2025. Densely connected parameter-efficient tuning for referring image segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Sehyun Hwang, Sohyun Lee, Hoyoung Kim, Minhyeon Oh, Jungseul Ok, and Suha Kwak. 2023. Active learning for semantic segmentation with multi-class label query. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. 2022. Pseudo-q: Generating pseudo language queries for visual grounding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Weitai Kang, Mengxue Qu, Yunchao Wei, and Yan Yan. 2024. Actress: Active retraining for semi-supervised visual grounding. *arXiv preprint arXiv:2407.03251*.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher D Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *the Association for Computational Linguistics (ACL)*.

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

Dong-Jin Kim, Jae Won Cho, Jinsoo Choi, Yunjae Jung, and In So Kweon. 2021. Single-modal entropy based active learning for visual question answering. In *British Machine Vision Conference (BMVC)*.

Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. 2023a. Shatter and gather: Learning referring image segmentation with text supervision. In *IEEE International Conference on Computer Vision (ICCV)*.

Hoyoung Kim, Sehyun Hwang, Suha Kwak, and Jungseul Ok. 2024a. Active label correction for semantic segmentation with foundation models. In *International Conference on Machine Learning (ICML)*.

Hoyoung Kim, Minhyeon Oh, Sehyun Hwang, Suha Kwak, and Jungseul Ok. 2023b. Adaptive superpixel for active learning in semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*.

Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. 2024b. Extending clip's image-text alignment to referring image segmentation. In *Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo,

and 1 others. 2023. Segment anything. In *IEEE International Conference on Computer Vision (ICCV)*.

Bo Li and Tommy Sonne Alstrøm. 2020. On uncertainty estimation in active learning for image segmentation. *arXiv preprint arXiv:2007.06364*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Xiao Lin and Devi Parikh. 2017. Active learning for visual question answering: An empirical study. *arXiv preprint arXiv:1711.01732*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Ting Liu, Zunnan Xu, Yue Hu, Liangtao Shi, Zhiqiang Wang, and Quanjun Yin. 2024. Mapper: Multimodal prior-guided parameter efficient tuning for referring expression comprehension. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mengyao Lyu, Jundong Zhou, Hui Chen, Yijie Huang, Dongdong Yu, Yaqian Li, Yandong Guo, Yuchen Guo, Liuyu Xiang, and Guiguang Ding. 2023. Box-level active detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. 2018. Cereals-cost-effective region-based active learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*.

Rafid Mahmood, Sanja Fidler, and Marc T Law. 2021. Low-budget active learning via wasserstein distance: An integer programming approach. In *International Conference on Learning Representations (ICLR)*.

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

Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. 2022. Towards robust and reproducible active learning using neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision (ECCV)*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2024. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*.

Niclas Popp, Dan Zhang, Jan Hendrik Metzen, Matthias Hein, and Lukas Schott. 2025. Foundation model-based data selection for dense prediction tasks. In *International Conference on Learning Representations (ICLR) Workshop on Foundation Models in the Wild*.

Practical Cryptography. English letter frequencies. http://www.practicalcryptography.com.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, and 1 others. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *IEEE International Conference on Computer Vision (ICCV)*.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Yawar Siddiqui, Julien Valentin, and Matthias Nießner. 2020. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiamu Sun, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. 2023. Refteacher: A strong baseline for semi-supervised referring expression comprehension. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. 2019. Bayesian generative active deep learning. In *International Conference on Machine Learning (ICML)*.

Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. 2012. Seeds: superpixels extracted via energy-driven sampling. In *European Conference on Computer Vision (ECCV)*.

AV Vladzymyrskyy, AV Gonchar, and V Yu Chernina. 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. In *IJCNN*.

XuDong Wang, Jingfeng Yang, and Trevor Darrell. 2024. Segment anything without supervision. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. 2024. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. 2024. Plug and play active learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research (TMLR)*.

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

Seonghoon Yu, Joonbeom Hong, Joonseok Lee, and Jeany Son. 2025. Latent expression generation for referring image segmentation and grounding. In *IEEE International Conference on Computer Vision (ICCV)*.

Seonghoon Yu, Ilchae Jung, Byeongju Han, Taeoh Kim, Yunho Kim, Dongyoon Wee, and Jeany Son. 2024a. A simple baseline with single-encoder for referring image segmentation. *arXiv preprint arXiv:2408.15521*.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. 2023. Zero-shot referring image segmentation with global-local context features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. 2024b. Pseudo-ris: Distinctive pseudo-supervision generation for referring image segmentation. In *European Conference on Computer Vision (ECCV)*.

Ying Zang, Chenglong Fu, Runlong Cao, Didi Zhu, Min Zhang, Wenjun Hu, Lanyun Zhu, and Tianrun Chen. 2024. Resmatch: Referring expression segmentation in a semi-supervised manner. *arXiv preprint arXiv:2402.05589*.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*.

Cost-efficient Active Learning for Referring Image Segmentation and Grounding
- *Appendix* -

Table of Contents

A. **Additional Experiments**
- A.1. Comparisons on Ambiguous Subset
- A.2. Comparisons with Additional Acquisition
- A.3. Adaptation Results on Medical Domains
- A.4. Auxiliary Region with Other Class
- A.5. Analysis on GPU Cost
- A.6. Comparisons with Visual-Similarity Baselines
- A.7. Validity of the Bit-Based Cost Model
- A.8. Analysis on Collected Expressions

B. **Additional Details**
- B.1. Theoretical Text Annotation Cost
- B.2. Acquisition Baselines
- B.3. User Study
- B.4. Dataset

C. **Additional Qualitative Results**
- C.1. Examples of User Study Results
- C.2. Examples of Auxiliary Region-Text Pairs
- C.3. Examples on Ambiguous Subset
- C.4. Examples of Selected Samples via Acquisitions
- C.5. Examples of Referred Region Ambiguity

A Additional Experiments

A.1 Comparisons on Ambiguous Subset

Our acquisition function focuses on ambiguous cases, where several objects similar to the target appear in the same image. To verify that it effectively improves the model's ability to distinguish the target from similar ones, we evaluate the RIS model trained under various acquisitions on an ambiguous subset in Tab. 5. We construct this subset from a val split of RefCOCO by selecting 100 heavily confused images that contain more than three objects of the same class (provided in the ground-truth labels). Our approach yields the best performance and among the smallest relative per-

Acquisition	# of Instances					
	2k		5k		10k	
	All	Subset	All	Subset	All	Subset
Random	38.46	32.93 (-14.4%)	55.64	48.80 (-12.3%)	65.23	60.89 (-6.7%)
Coreset	33.54	30.51 (-9.0%)	53.77	43.82 (-18.5%)	64.33	54.73 (-14.9%)
Pixel-entropy	39.25	35.74 (-8.9%)	56.62	48.04 (-15.2%)	65.48	60.56 (-7.5%)
Ours	49.23	**47.03** (-4.5%)	62.91	**58.49** (-7.0%)	68.96	**64.00** (-7.2%)

Table 5: Comparisons with several acquisition baselines on an ambiguous subset of 100 highly complex images, each containing more than three objects of the same class.

Acquisition	# of Instances		
	2k	5k	10k
PPAL (Yang et al., 2024)	36.77	55.59	65.98
Learning Loss (Yoo and Kweon, 2019)	39.79	54.82	64.03
Referred Region Ambiguity (ours)	**49.14**	**61.98**	**67.94**

Table 6: Comparisons with additional acquisition baselines for a RIS task on the validation set of RefCOCO.

Acquisition	QaTa-COV19				MosMedData+			
	# of instances				# of instances			
	0k†	0.1k	0.5k	1k	0k†	0.1k	0.3k	0.5k
Random	12.23	48.72	60.47	64.09	4.52	39.27	46.89	49.52
Coreset		46.96	61.01	64.28		37.17	45.60	48.57
Pixel-entropy		48.95	59.16	64.22		37.10	46.22	47.54
Ours		**49.79**	**62.91**	**65.40**		**39.68**	**47.12**	**50.26**

Table 7: Adaptation results on medical domains compared with several acquisitions. † denotes the zero-shot results of the RIS model fully-trained on RefCOCO.

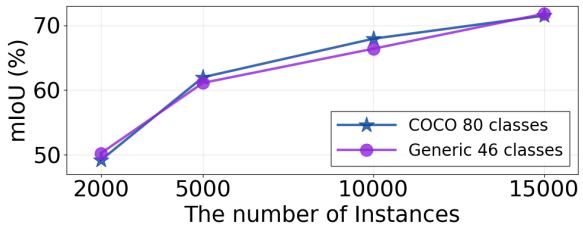


Figure 13: Comparison with different class sets for the auxiliary region extraction (Sec. 3.2 in the manuscript) on RefCOCO dataset for a RIS task. Each class set is prompted into Grounded-SAM to extract auxiliary regions. The generic 46-class set is shown in Tab. 8.

formance drops (%) on this subset, highlighting that it indeed enhances the discriminative ability in complex scenes. The qualitative results of predictions for each method on this set are visualized in Fig. 19.

A.2 Comparisons with Additional Acquisition

In Tab. 6, we compare our referred region ambiguity with additional acquisition baselines, including PPAL (Yang et al., 2024) and Learning Loss (Yoo and Kweon, 2019), on the RefCOCO validation set for a RIS task. To isolate the effect of acquisition from annotation tools, we evaluate all methods using the number of acquired samples as the budget. Our acquisition achieves the best mIoU, consistent with the results in Fig. 7 of the main manuscript. These results further demonstrate the effectiveness of our acquisition in improving the fine-grained discriminative ability required for VG.

Generic 46-class set
person, animal, pet, bird, mammal, sea creature, insect, vehicle, car, truck, motorcycle, bicycle, aircraft, watercraft, furniture, appliance, electronics, kitchenware, decoration, clothing, footwear, accessories, bag, jewelry, food, fruit, vegetable, meal, snack, beverage, ingredient, tool, utensil, sports equipment, musical instrument, device, outdoor object, sign, light, plant, building element, object, container, item, product, material

Table 8: List of a generic 46-class set, constructed by abbreviating Object365 categories using ChatGPT.

A.3 Adaptation Results on Medical Domains

To examine how well our AL framework efficiently supports the domain adaptation, we fine-tune the RIS model (pretrained on RefCOCO) on medical domains with various acquisitions in Tab. 7. We conduct experiments on two medical referring image segmentation datasets: QaTa-COV19 (Degerli et al., 2022) (5,716 train samples) and MosMedData+ (Vladzymyrskyy et al., 2020) (2,183 train samples). Due to a large domain gap (leading to poor zero-shot results), other methods may be less reliable, making random selection a competitive baseline. Nevertheless, our acquisition exhibits faster adaptation than all the compared methods, suggesting its effectiveness in the data-sparse domains with few samples.

A.4 Auxiliary Region with Other Class

In *Generation of Auxiliary Region-Text Pairs* (Sec. 3.2 of the main manuscript), we prompt COCO (Lin et al., 2014) 80 classes into Grounded-SAM (Ren et al., 2024) to extract auxiliary regions across all unlabeled images. To further validate the

Types	Generation of Auxiliary Pairs	Annotation	Total Cost	GPU Memory
GPU overhead (w/ ours)	16 hours	844 hours	$120.4	8.5GB
Human labor (w/ ours)	-	844 hours	$6,752	-
Human labor (w/o ours)	-	2,287 hours	$18,296	-

Table 9: Comparisons of total annotation cost in U.S. dollars, including GPU overhead and human labor in our approach, compared with the standard baseline with naïve labeling and random selection. GPU cost is estimated using an RTX 3090 at $0.14/hour on Vast.ai, and human labor cost is calculated assuming an hourly wage of $8/hour.

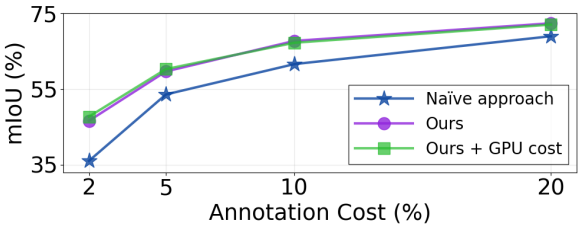


Figure 14: Comparisons under a fixed U.S. dollar budget that accounts for the GPU overhead in our approach. The cost estimates are summarized in Tab. 9.

generality of our method beyond COCO 80 categories, we conduct auxiliary region extraction with a more generic 46-class set and report the mIoU results under varying numbers of labeled instances on RefCOCO dataset (averaged over val, testA, and testB results for a RIS task) in Fig. 13. To construct the generic 46-class vocabulary, we aggregate the fine-grained Object365 (Shao et al., 2019) categories into 46 broader superclasses using ChatGPT. The performance gap between a COCO 80-class set and a generic 46-class set is minimal, indicating that our method is robust to the choice of object vocabulary used for generating auxiliary regions.

A.5 Analysis on GPU Cost

The proposed AL framework introduces additional GPU overhead beyond the standard model training in two parts: (1) generating auxiliary region-text pairs using Grounded-SAM (Ren et al., 2024) and ViP-LLaVA (Cai et al., 2024), and (2) supporting annotation by running ViP-LLaVA in the text labeling interface for text annotation and pre-extracting SAM (Kirillov et al., 2023) masks offline for the region labeling. In this subsection, we quantify these additional GPU costs in U.S. dollars (Tab. 9), compare them with human labor costs, and report experiments that explicitly account for resulting GPU overhead from both components (Fig. 14).

GPU Overhead in Generation of Auxiliary Pairs. To generate auxiliary pairs, we use a single RTX 3090 GPU rented from Vast.ai at $0.14/hour. We run Grounded-SAM for 4 hours to produce auxiliary regions and ViP-LLaVA for 12 hours to generate auxiliary expressions, resulting in a total of 16 GPU hours. This corresponds to an overall cost of $2.24 (16 hours $\times$ $0.14/hour).

GPU Overhead in Labeling Interface. For text annotation, running ViP-LLaVA in the labeling interface requires 840 GPU hours on a single RTX 3090, computed from the per-sample latency in the user study (Tab. 2) of the main manuscript over the entire 42,404 RefCOCO samples. At $0.14/hour on Vast.ai, this amounts to $117.6. For region annotation, we pre-extract SAM masks offline, which takes 4 GPU hours on the same GPU and costs $0.56. In total, the labeling interface requires 844 GPU hours, corresponding to a total GPU cost of $118.16. In addition, our labeling tool still requires human annotation, assuming an hourly wage of $8, the 844 hours of human labor correspond to $6,752, as reported in Tab. 9. We also report the incurred GPU memory (GB) of ViP-LLaVA-7B on an NVIDIA RTX 3090 GPU in Tab. 9.

Compared to the Human Labor Costs. To compare these costs incurred in our approach with human manual labeling costs, we estimate the total annotation time for all 42,404 RefCOCO samples, based on the per-sample latency reported in the user study (Tab. 2) of the main manuscript. The naïve polygon drawing for the region annotation requires 761 hours, while manually typing full sentences for the text annotation demands 1,526 hours. In total, this amounts to 2,287 hours of human labor. Assuming an hourly wage of $8, the total human labor cost is $18,296, which is substantially higher than the GPU cost incurred by our labeling interface and auxiliary region-text generation. In Fig. 14, we further report results under a fixed U.S. dollar budget (%) that accounts for GPU costs, comparing naïve labeling with random acquisition against our labeling with our acquisition. Despite the additional GPU overhead in our approach (Tab. 9), the results in Fig. 14 show that our method remains more cost-

Acquisition	# of Instances		
	2k	5k	10k
Random	37.79	54.36	63.87
DINOv2 region similarity	46.20	56.00	65.36
Max-region selection	46.32	59.99	66.86
Referred Region Ambiguity (ours)	**49.14**	**61.98**	**67.94**

Table 10: Comparisons with acquisition baselines that rely only on visual similarity or object density, for a RIS task on the validation set of RefCOCO.

Interface	Modeled Cost		Measured Time	
	Bits	Reduction	Time (s)	Reduction
Naïve typing	6,493,300	0.0%	45.82	0.0%
Sentence selection	2,814,811	56.7%	21.10	53.9%
Word selection (ours)	4,238,677	34.7%	28.61	37.6%

Table 11: Validation of the bit-based text annotation cost. We compare the reduction in the modeled bit cost with the reduction in the annotation time measured in user studies, both relative to naïve typing.

effective under the same budget. This is because the additional GPU cost is minimal compared with the human labor cost required for large-scale manual annotation.

A.6 Comparisons with Visual-Similarity Baselines

To examine whether the gains of our acquisition arise primarily from selecting images that contain many visually similar objects, we compare referred region ambiguity with two additional baselines that rely only on visual signals. For *DINOv2 region similarity*, we extract DINOv2 (Oquab et al., 2024) features for all candidate regions in an image and use their average pairwise cosine similarity as the acquisition score. For *max-region selection*, we rank images by the number of regions proposed by Grounded-SAM (Ren et al., 2024), prioritizing images with more candidate objects. Tab. 10 reports the results on the RefCOCO validation set for a RIS task, following the protocol of Appendix A.2. Both baselines outperform random selection, suggesting that visual similarity and object density are useful acquisition signals. However, they consistently underperform referred region ambiguity across all annotation budgets. These results indicate that our gains cannot be explained solely by selecting images with many visually similar objects. Rather, our criterion captures language-conditioned, model-specific uncertainty, *i.e.*, cases in which the current grounding model cannot reliably distinguish the referred target from similar candidates given the expression.

A.7 Validity of the Bit-Based Cost Model

To further examine the validity of the bit-based text annotation cost (Eq. (3) in Sec. 4 of the main manuscript), we conduct an additional user study for the sentence-selection interface in Tab. 3 on Amazon Mechanical Turk. Twenty annotators each label 10 samples, yielding 200 annotations in total. We compare the reduction in the modeled bit cost with the reduction in the measured annotation time, both relative to naïve typing. As shown in Tab. 11, the modeled bit-cost reduction of sentence selection closely aligns with the measured time reduction (56.7% vs. 53.9%), and a similar agreement is observed for our word-selection interface (34.7% vs. 37.6%). Although the model cannot capture every aspect of human effort, these consistent trends across both interfaces provide additional evidence that it is a reasonable proxy for the relative annotation cost.

A.8 Analysis on Collected Expressions

Beyond the annotation time and similarity metrics in Tab. 2, we further analyze the referring expressions collected with naïve typing and with our interface in the user study. Tab. 12 reports the following statistics. We use ChatGPT to categorize the words in each expression as attributes, spatial relations, or object categories, and report their proportions. We measure diversity using the Vendi Score (Friedman and Dieng, 2023), which quantifies the semantic diversity of the collected expressions based on LLM embeddings. To assess ambiguity, we conduct a paired comparison in which ChatGPT identifies the more ambiguous expression between those produced by naïve typing and our interface.

Tool	Avg. Len.	Word Type			Diversity ↑	Ambiguity ↓	Quality ↑	Refer. Acc. ↑
		Attribute	Spatial	Category				
Naïve typing	12.0	23.6%	15.8%	9.8%	**20.33**	55.5%	3.88	34.70%
Ours	6.2	23.2%	17.4%	17.4%	17.71	**44.5%**	**3.95**	**65.30%**

Table 12: Language-side analysis of the referring expressions collected with naïve typing and with our interface. Avg. Len. is the average number of words. Word Type is the proportion of attribute, spatial-relation, and object-category words. Diversity is the Vendi Score (Friedman and Dieng, 2023). Ambiguity is the ratio of being judged as the more ambiguous expression in a paired comparison. Quality is a human-rated five-point score from 90 annotators. Refer. Acc. is the ratio of being judged as the expression that more accurately refers to the target.

For human-rated quality, we conduct an additional user study with 90 annotators from Amazon Mechanical Turk, where each annotator evaluates 20 expressions and rates their quality on a five-point scale. To assess referring accuracy, we ask ChatGPT to identify which expression more accurately refers to the target object.

Naïve typing yields a higher diversity score, likely because its expressions are longer and contain more concepts. However, our manual inspection shows that some of these annotations describe the overall image rather than the target region. For example, one annotation states: *"Three ladies are seated close together on a sofa; one is using headphones to unwind, another is wearing a robe and face mask and using a laptop to examine her fingernails, and a fourth figure is depicted as a blue, overlaid silhouette sitting on the right."* This expression describes the overall scene and includes inconsistent details rather than clearly identifying the target region. Such errors may increase diversity but also reduce clarity and increase ambiguity. In contrast, our interface produces clearer and more target-focused expressions, as reflected in the lower ambiguity, higher human-rated quality, and substantially higher referring accuracy.

B Additional Details

B.1 Theoretical Text Annotation Cost

In our text annotation cost measurement (Eq. (3) in the main manuscript, rewritten below for convenience), we rely on two types of probabilities: (1) the probability p_k that a user-desired word appears among the suggested candidates in the annotation interface (*i.e.*, top-k coverage probability), and (2) the conditional probability of a letter given the previous two letters, $p_l(l_i|l_{i-1}, l_{i-2})$ (*i.e.*, trigram probability). In this subsection, we provide detailed statistics and empirical measurements for these two probabilities.

$$C_{\text{word}} = \underbrace{\log_2(k+1)}_{\text{Click a word or typing option}} + \underbrace{(1-p_k)C_{\text{type}}}_{\text{Type a word manually}},$$

$$\text{where } C_{\text{type}} = -\sum_{i=1}^{n} \log_2 p_l(l_i|l_{i-1}, l_{i-2}).$$

Top-k Suggestions in Text Annotation Tool. To estimate top-k coverage probability p_k, we measure the top-k accuracy of the k-word suggestions produced by our text annotation tool on the RefCOCO dataset. At each step of the annotation process, we mark the prediction as correct if the ground-truth next word appears within the top-k suggested candidates. The ratio of correct predictions across all steps yields the observed top-k accuracy, which we use as p_k. As shown in Tab. 13, p_k naturally increases as k grows. However, the overall measured cost still increases because selecting from a larger candidate set incurs a higher option cost, represented by the $\log_2(k+1)$ term. We choose $k = 5$ as it provides a more intuitive word selection while achieving the lowest measured annotation cost.

Conditional Probability of Letters. For the trigram conditional probability of a letter, $p_l(l_i|l_{i-1}, l_{i-2})$ in Eq. (3) of text cost measurement, we use letter-level 3-gram frequencies measured from approximately 4.5 billion characters of English text (we use the English letter frequency data presented in (Practical Cryptography)). For the first and second letters of a sequence, we use 1-gram and 2-gram frequencies, respectively. Subsets of these letter frequencies are shown in Tab. 14.

B.2 Acquisition Baselines

We conduct acquisition comparisons with pixel entropy (Mackowiak et al., 2018) and coreset (Sener and Savarese, 2018) in Fig. 7 (RIS) and Fig. 8 (REC) of the main manuscript. In this subsection, we provide a detailed implementation for each of the compared baselines.

Metric	Top-k candidates			
	3	**5** (ours)	10	20
Accuracy (%)	56.48	61.75	67.10	**71.21**
Measured Cost over Dataset	4,274,759	**4,238,677**	4,324,874	4,524,190

Table 13: Text annotation costs and accuracy across varying k suggested word candidates in our text annotation interface for estimating p_k, which represents the top-k accuracy of including the ground-truth next word among the suggested k candidates over the RefCOCO dataset.

	1-gram		**2-gram**		**3-gram**	
Rank	Letters	Probability	Letters	Probability	Letters	Probability
1	E	0.1210	TH	0.0272	THE	0.0181
2	T	0.0894	HE	0.0234	AND	0.0073
3	A	0.0855	IN	0.0204	ING	0.0072
4	O	0.0747	ER	0.0179	ENT	0.0042
5	I	0.0733	AN	0.0162	ION	0.0042
6	N	0.0717	RE	0.0142	HER	0.0036
7	S	0.0673	ES	0.0133	FOR	0.0034
8	R	0.0633	ON	0.0132	THA	0.0033
9	H	0.0496	ST	0.0126	NTH	0.0033
10	L	0.0421	NT	0.0118	INT	0.0032

Table 14: Top-10 most frequent letter n-grams with their probabilities. We use the English letter frequency data presented in (Practical Cryptography).

Pixel Entropy. This method is a conventional acquisition function widely used in active learning for segmentation (Mackowiak et al., 2018; Li and Alstrøm, 2020; Siddiqui et al., 2020). It quantifies the informativeness of an auxiliary region-text pair by measuring how uncertain the model's prediction is at the pixel level, and it selects region-level samples for human annotations instead of image-level in ours. Pixel entropy outputs high informativeness when the model's confidence map is uniformly spread across the whole image with moderate values (*e.g.*, probability near 0.5). Conversely, if the model confidently activates the whole image with the lowest or highest values (*i.e.*, probability near 0.0 or 1.0), it outputs low informativeness. We adapt this method for both RIS and REC models, as follows.

1) RIS model: Given a specific auxiliary pair $(r^{\text{aux}}, t^{\text{aux}})$, we obtain a confidence map (*i.e.*, a final logit scoring map before thresholding), $p^i = \theta(I, t^{\text{aux}}) \in \mathbb{R}^{h \times w}$. We then apply a sigmoid function to convert these logits into a pixel-level probability map, as follows: $\hat{p} = \sigma(p) \in [0,1]^{h \times w}$. The binary entropy $\hat{H}_{(x,y)}$ at each pixel (x, y) is defined as:

$$\begin{aligned}\hat{H}_{(x,y)} = &- \hat{p}_{(x,y)} \log(\hat{p}_{(x,y)}) \\ &- (1 - \hat{p}_{(x,y)}) \log(1 - \hat{p}_{(x,y)}),\end{aligned} \quad (4)$$

where $\hat{H}_{(x,y)}$ is the binary entropy at pixel coordinate (x, y) conditioned on a given auxiliary caption t^{aux}. The final informativeness score $E(r^{\text{aux}}, t^{\text{aux}})$ for an auxiliary pair is obtained by averaging this pixel-level entropy $\hat{H}_{(x,y)}$ across all $h \times w$ pixels, as follows:

$$E(r^{\text{aux}}, t^{\text{aux}}) = \frac{1}{h \times w} \sum_{x=1}^{h} \sum_{y=1}^{w} \hat{H}^i_{(x,y)}. \quad (5)$$

Based on this $E(r^{\text{aux}}, t^{\text{aux}})$ score computed for all auxiliary pairs, we select the top-scoring auxiliary region-text pairs and annotate them at the region-level sample.

2) REC model: Unlike RIS models that directly generate segmentation masks from a probability map, REC models predict bounding boxes by passing only the [CLS] token through an MLP layer, making pixel-level entropy inapplicable. Instead, we leverage the attention map $A \in \mathbb{R}^{h \times w}$ extracted from the final decoder layer, conditioned on the given auxiliary caption t^{aux} and image I. We compute the normalized entropy H_n of this attention map and then average across all N auxiliary captions, as follows:

$$H_n(A) = -\frac{\sum_{x=1}^{h} \sum_{y=1}^{w} A_{(x,y)} \log(A_{(x,y)})}{\log(h \times w)}, \quad (6)$$

where $A_{(x,y)}$ is an attention score at a pixel coordinate (x, y). We use this $H_n(A)$ as a pixel-entropy score for each auxiliary pair in a REC model. We then select and annotate the auxiliary pairs with the highest $H_n(A)$ scores sequentially.

Coreset. This method aims to select diverse samples from an unlabeled pool to maximize coverage of acquired samples. To achieve this, it first extracts feature representations for all auxiliary pairs in the unlabeled auxiliary region-text pair set $\mathcal{D}_\mathcal{U}$ and then iteratively selects the auxiliary pair whose feature is farthest from the features of all currently labeled region-text pair set $\mathcal{D}_\mathcal{L}$.

Formally, given a specific auxiliary caption t^{aux} with its image (for a labeled case, we use a ground-truth expression instead of an auxiliary caption), we extract the feature representation of a given auxiliary caption from the model encoder θ_{enc}, as follows:

$$F = \theta_{\text{enc}}(I, t^{\text{aux}}) \in \mathbb{R}^{h \times w \times d}, \tag{7}$$

where h, w are the size of the feature map and d is a feature dimension. The resulting feature map F is then aggregated via spatial pooling and ℓ_2-normalized to obtain the final image feature vector $\hat{Z}$:

$$\hat{Z} = \frac{Z}{\|Z\|}, \text{ where } Z = \frac{\sum_{x=1}^{h} \sum_{y=1}^{w} F_{(x,y)}}{h \times w} \in \mathbb{R}^d, \tag{8}$$

and $F_{(x,y)}$ denotes the feature at a pixel coordinate (x, y). This normalized feature vector $\hat{Z}$ represents the auxiliary pair feature and is used for distance calculation. For each unlabeled auxiliary pair, we compute its minimum distance over all labeled region-text pairs in the acquired pair set $\mathcal{D}_\mathcal{L}$:

$$d_j = \min_{k \in \mathcal{D}_\mathcal{L}} \|\hat{Z}_j - \hat{Z}_k\|_2, \tag{9}$$

where k denotes the index for the labeled pairs. We select the auxiliary pair with the maximum distance to the current labeled set:

$$m^* = \arg\max_{j \in \mathcal{D}_\mathcal{U}} d_j, \tag{10}$$

where m^* is a selected index over all auxiliary pairs in the unlabeled pairs set $\mathcal{D}_\mathcal{U}$. We add it to the labeled set, $\mathcal{D}_\mathcal{L} \leftarrow \mathcal{D}_\mathcal{L} \cup \{(r^{\text{aux}}_{m^*}, t^{\text{aux}}_{m^*})\}$ with its annotations and an image. We then update the distances for all remaining auxiliary pairs: $d_j \leftarrow \min(d_j, \|\hat{Z}_j - \hat{Z}_{m^*}\|), \forall j \in \{1, \ldots, |\mathcal{D}_\mathcal{U}|\} \backslash \{m^*\}$. This process iteratively continues until the round budget is exhausted.

B.3 User Study

To verify the efficiency of the proposed annotation interface, we conduct a user study with 30 annotators from Amazon Mechanical Turk in Tab. 2 of the main manuscript. In this subsection, we provide the details of the user study. Each annotator is required to complete annotations for 10 images, following the instructions in Fig. 15; the instructions also include example videos for easy understanding. We record the time required to annotate both the mask and referring expression and save the annotation results (Fig. 16) to compare the quality with the ground truth labels. Besides our proposed interface, we also conduct the same user study for a naïve annotation (*i.e.*, polygon drawing for masks and manual typing for expressions) to evaluate the benefits of our proposed tool. The example videos using our labeling interface are provided in our GitHub repository.

B.4 Dataset

RefCOCO. RefCOCO (Yu et al., 2016) provides images paired with masks or boxes and their corresponding referring text. This is created using the ReferItGame (Kazemzadeh et al., 2014), where one player writes a textual description for a specific instance, and a second player locates the correct object based on that description. The dataset contains 142,210 expressions referring to 50,000 instances across 19,994 images. These instances are divided into the following splits: 42,404 for a training set, 3,811 for a validation set, 1,975 for a testA set, and 1,810 for a testB set. The expressions in RefCOCO are typically concise and often include directional instructions, such as "*left*" or "*right*", to disambiguate similar objects.

RefCOCO+. RefCOCO+ (Yu et al., 2016) is constructed using the same ReferItGame (Kazemzadeh et al., 2014) method as RefCOCO. The difference in textual form is that it explicitly prohibits the use of positional terms. This constraint forces the expressions to be based primarily on object attributes, such as clothing color or object size. This distinction allows researchers to evaluate a model's performance on attribute comprehension, separate from its understanding of spatial cues. It includes

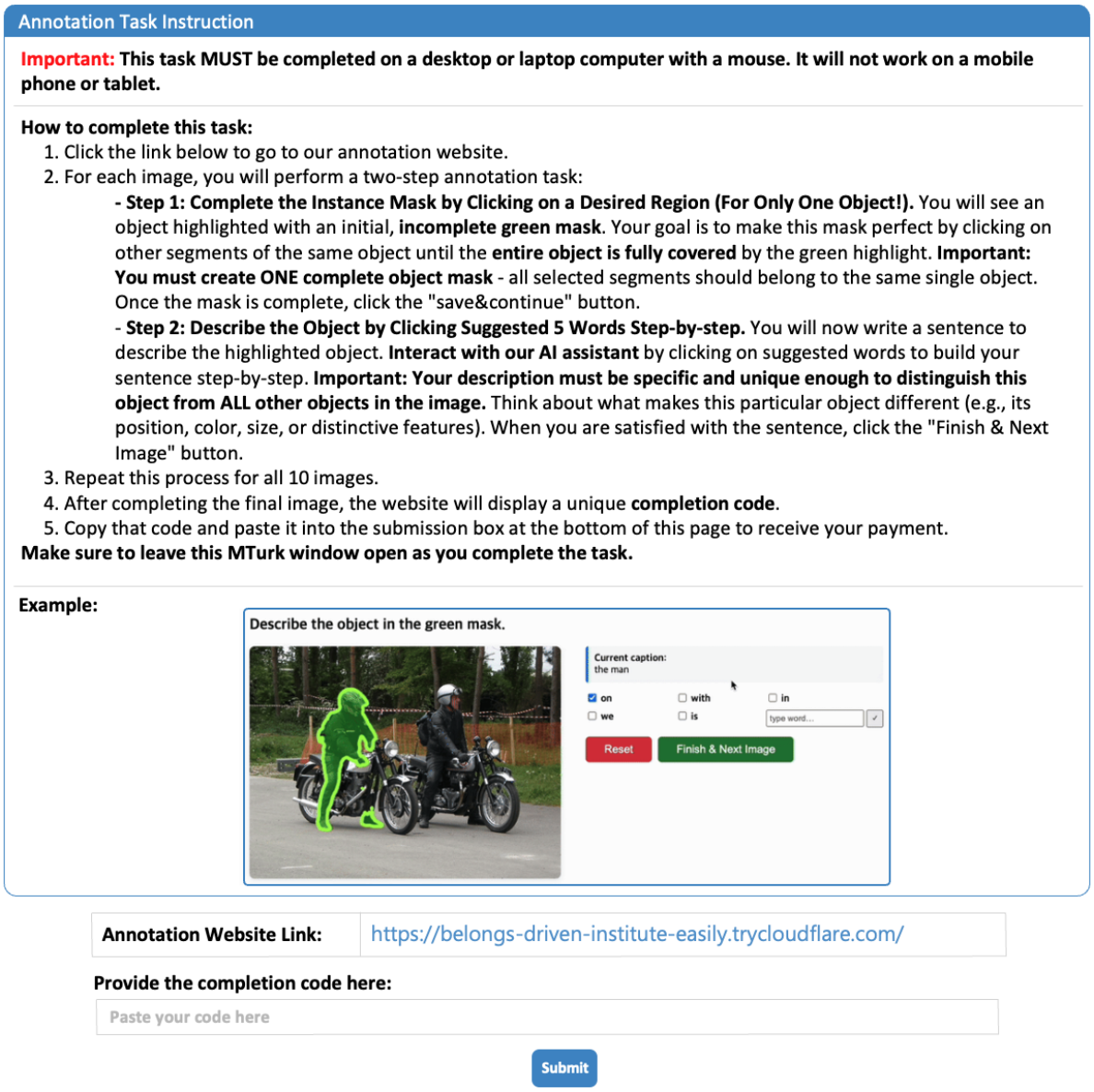


Figure 15: Annotation task instructions for the user study on Amazon Mechanical Turk.

141,564 expressions for 49,856 instances across 19,992 images. These instances are divided into 42,278 for a train set, 3,805 for a val set, 1,975 for a testA set, and 1,798 for a testB set, following a similar split structure to RefCOCO.

RefCOCOg. RefCOCOg (Nagaraja et al., 2016) is characterized by much longer expressions, averaging 8.43 words per expression, compared to 3.61 and 3.53 words for RefCOCO and RefCOCO+, respectively. Therefore, it demands a model's ability for more complex and contextual comprehension and is commonly used to evaluate multi-modal reasoning under longer linguistic context. Compared to RefCOCO and RefCOCO+, which are built on ReferItGame (Kazemzadeh et al., 2014), RefCOCOg is collected via Amazon Mechanical Turk. This dataset includes 95,010 expressions for 49,822 object instances across 25,799 images. These instances are separated into 42,226 for a train split, 2,573 for a validation split, and 5,023 for a test split.

C Additional Qualitative Results

C.1 Examples of User Study Results

We conduct a user study with 30 annotators from Amazon Mechanical Turk in Tab. 2 of the main manuscript. In Fig. 16, we provide qualitative examples of those results compared with the naïve labeling tool (*i.e.*, polygon drawing for masks and manual typing for expressions). The masks an-

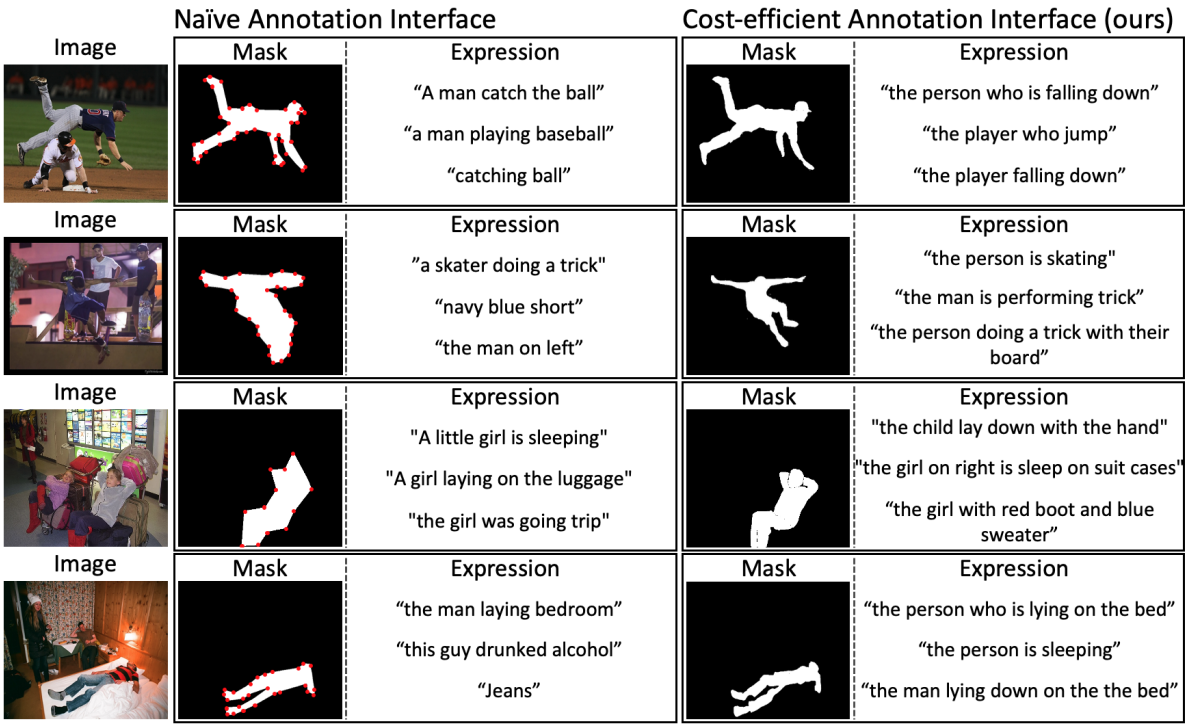


Figure 16: Examples of user study results with our cost-efficient annotation interface compared to the naïve annotation interface (*i.e.*, polygon drawing for masks and manual typing for expressions).



Figure 17: Examples of the generated auxiliary region-text pairs. The paired region-text are denoted by matching colors.

notated via our interface are more accurate than those drawn manually, since polygon drawing is labor-intensive and struggles with complex object boundaries. Our mask tool leverages precise pre-extracted SAM masks, enabling high-quality annotations with faster times. Similarly, the expressions annotated through our interface accurately describe the referred regions. This is confirmed by the higher Sentence-BERT similarity scores compared to manual typing in Tab. 2 of the main manuscript. This is because, at each step, the decoding process of the multi-modal language model (*i.e.*, ViP-LLaVA (Cai et al., 2024)) suggests contextually appropriate words relevant to the target region, guiding annotators to concise and accurate descriptions.

C.2 Examples of Auxiliary Region-Text Pairs

We provide qualitative examples of the generated auxiliary region-text pairs (Sec. 3.2 of the main

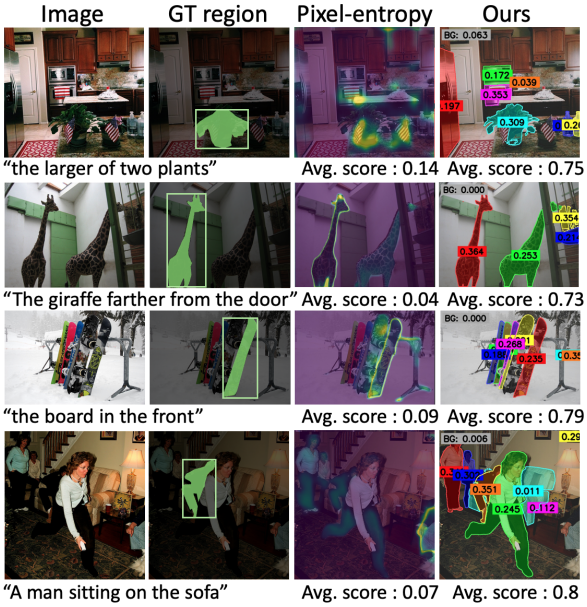


Figure 18: Additional qualitative analysis on the process of computing the informativeness score compared to the pixel-entropy.

manuscript) in Fig. 17, where each region-text pair is highlighted using the same colors. Although some pairs contain minor noise, the auxiliary regions accurately localize the objects, and the corresponding captions provide clear and detailed referring expressions for each region, as quantified in Tab. 4.

C.3 Examples on Ambiguous Subset

In Tab. 5 of Appendix A.1, we conduct comparisons with several acquisition baselines (*i.e.*, random, coreset, pixel-entropy, and ours) on the curated ambiguous test subset. In this subsection, we present their qualitative results from this ambiguous subset in Fig. 19. Compared with other acquisitions, the RIS model trained under our acquisition performs better at distinguishing the target from visually similar distractors. This result suggests that the proposed referred region ambiguity is effective in improving the model's discriminative ability in challenging ambiguous cases.

C.4 Examples of Selected Samples via Acquisitions

In Fig. 20, we present additional examples of selected samples via each acquisition. Compared with other approaches (*i.e.*, pixel-entropy and coreset), our referred region ambiguity tends to prioritize images containing complex scenes with many objects, leading to more informative selections.

C.5 Examples of Referred Region Ambiguity

We show additional examples of the process for computing the informativeness score in our referred region ambiguity, compared with the pixel-entropy in Fig. 18. Although the images contain visually similar objects, pixel-entropy assigns low informativeness because it averages entropy over all pixels. In contrast, our acquisition function yields high scores by operating at the region level, effectively capturing the complexity of these scenes.

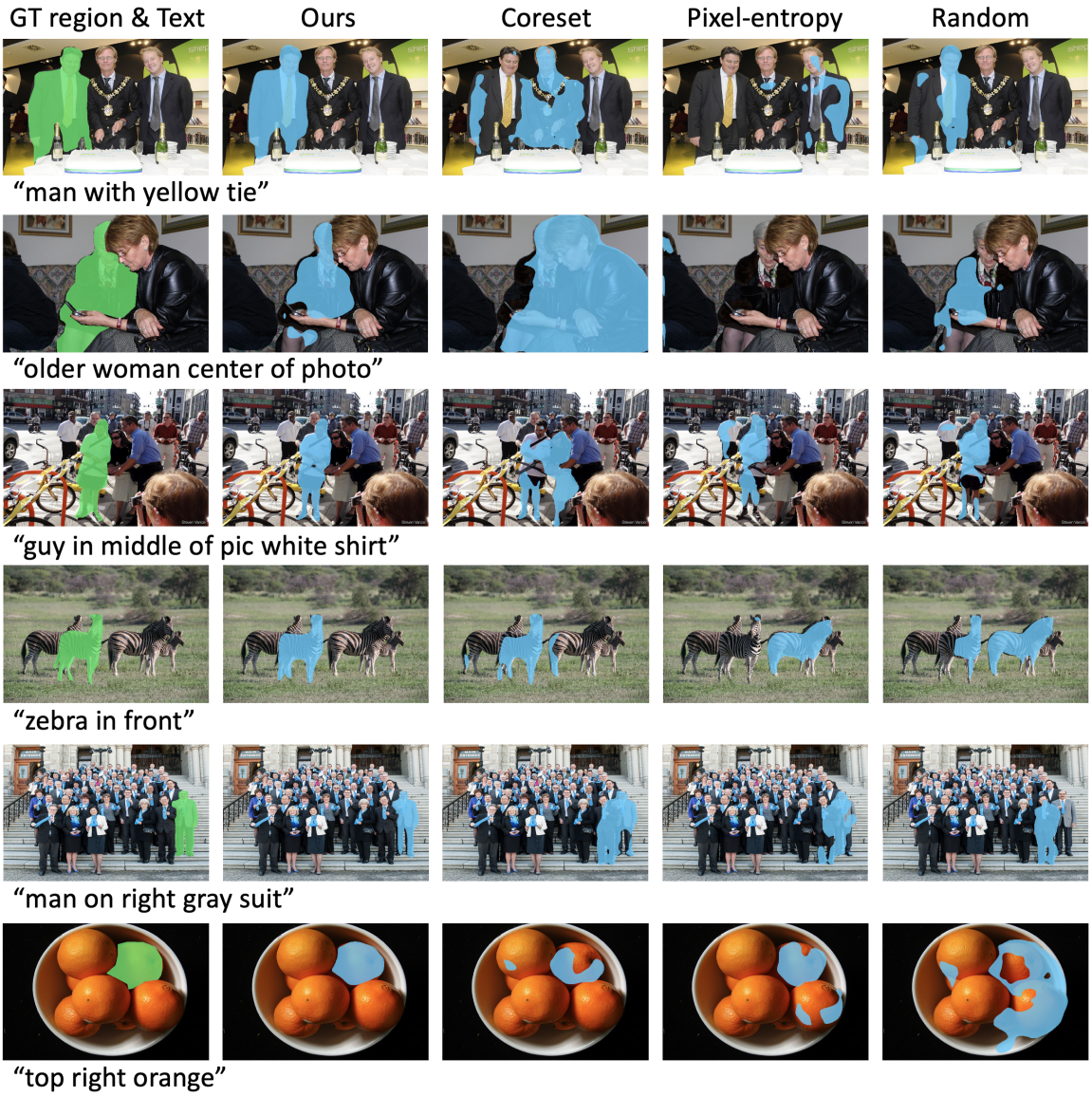


Figure 19: Qualitative examples from the ambiguous subset compared with the RIS models trained under different acquisition strategies.



(a) Referred region ambiguity (b) Pixel-entropy (c) Coreset

Figure 20: Additional examples of selected samples via acquisitions. The paired GT region-text are denoted by matching colors.