IntRec: Intent-based Retrieval with Contrastive Refinement

Pourya Shamsolmoali
Dept. Computer Science
University of York
York, United Kingdom

Masoumeh Zareapoor
Dept. Computer Science
Shanghai Jiao Tong University
Shanghai, China

Eric Granger
LIVIA, Dept. of Systems Engineering
École de technologie supérieure
Montreal, Canada

Yue Lu
School of Communication and Electronic Eng.
East China Normal University
Shanghai, China

ABSTRACT

Retrieving user-specified objects from complex scenes remains a challenging task, especially when queries are ambiguous or involve multiple similar objects. Existing open-vocabulary detectors operate in a one-shot manner, lacking the ability to refine predictions based on user feedback. To address this, we propose IntRec, an interactive object retrieval framework that refines predictions based on user feedback. At its core is an Intent State (IS) that maintains dual memory sets for positive anchors (confirmed cues) and negative constraints (rejected hypotheses). A contrastive alignment function ranks candidate objects by maximizing similarity to positive cues while penalizing rejected ones, enabling fine-grained disambiguation in cluttered scenes. Our interactive framework provides substantial improvements in retrieval accuracy without additional supervision. On LVIS, IntRec achieves 35.4 AP, outperforming OVMR, CoDet, and CAKE by +2.3, +3.7, and +0.5, respectively. On the challenging LVIS-Ambiguous benchmark, it improves performance by +7.9 AP over its one-shot baseline after a single corrective feedback, with less than 30 ms of added latency per interaction.

KEYWORDS

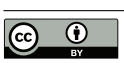
Interactive learning, Contrastive alignment, Object detection

ACM Reference Format:
Pourya Shamsolmoali, Masoumeh Zareapoor, Eric Granger, and Yue Lu. 2026. IntRec: Intent-based Retrieval with Contrastive Refinement. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/AGOG8131

1 INTRODUCTION

Modern visual systems are increasingly expected to understand complex user intent to retrieve specific objects within open-world environments [3, 18, 37]. This capability, termed object retrieval, requires the precise localization of a target based on varying linguistic or visual inputs [10, 22, 24]. Unlike traditional object detection, which predicts instances from a fixed category set [18], object retrieval focuses on the specific instance that aligns with a user's intent. This distinction is foundational for interactive applications such as human-robot collaboration, AR/VR assistance, and advanced visual search. Recent progress in open-vocabulary detection [4, 8, 41] and visual grounding [15] have enabled models to transcend fixed label sets by aligning image and text embeddings in a shared semantic space. However, these models operate under a one-shot retrieval design: a single query is matched to candidate regions, and the top-scoring region is returned as the prediction. This design is inherently limited. When queries are ambiguous or fine-grained (e.g., "the smaller umbrella with a floral pattern"), one-shot approaches produce incorrect or inconsistent predictions [8, 22, 40]. Recent efforts like OVMR [22] and MMOVD [10], attempt to mitigate this ambiguity by incorporating multimodal cues, yet they still struggle in cluttered scenes where user prompts are vague or underspecified. Formally, given a single text query T and a set of candidate image regions $R = \{r_1, \ldots, r_M\}$, standard open-vocabulary detectors identify the most likely target region by computing the similarity between the query embedding $E_T(T)$ and each region feature r_j. The region with the highest similarity score is selected as the predicted object, $b^\star = f_{\text{standard}}(T, R) = \text{argmax}\left(\text{sim}(E_T(T), r_j)\right)$.

This retrieval function is stateless, its output depends only on the query embedding and region features, with no mechanism to incorporate user feedback. In practice, these systems act as one-shot matchers that lack the temporal or logic-based depth to resolve ambiguity, a shortcoming that mirrors distribution misalignment issues found in generative diffusion tasks [32]. This failure is most evident in cluttered scenes containing distractors (i.e., competing objects that are visually similar to the target). For example, a prompt like "the smaller red car" may result in multiple visually similar objects, making it difficult for the model to determine which object the user actually intended. To address this limitation, we propose Intent-based Retrieval (IntRec), an interactive and stateful framework for open-world object localization. At the core of IntRec is the Intent State (IS), a memory structure that accumulates both positive anchors (user-confirmed cues) and negative constraints (explicit rejections). A contrastive ranking function then evaluates candidate regions by maximizing similarity to positive anchors while penalizing similarity to negative ones. This mechanism enables fine-grained disambiguation, allowing the model to distinguish between highly similar objects after even a single corrective feedback. As illustrated in Figure 1, IntRec operates through a multi-turn interaction loop, where user feedback updates the intent state. Through this process, the model progressively aligns with the user intent, resolving fine-grained ambiguities that one-shot detectors cannot



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/AGOG8131

handle. We evaluate IntRec on large-scale open-vocabulary detection benchmarks, comparing against state-of-the-art baselines using AP for overall performance and Rare-AP for novel object detection.

- We formulate object retrieval as an interactive intent refinement problem, addressing the ambiguity limitations of open-vocabulary detectors.
- We propose an Intent State module that accumulates both positive anchors and negative constraints from user feedback. A contrastive ranking function uses this state to refine retrieval and disambiguate fine-grained targets.
- Extensive experiments show our model consistently outperforms state-of-the-art methods.

2 RELATED WORK

2.1 Open-Vocabulary Detection and Grounding

Traditional object detectors rely on a fixed set of predefined labels, which limits their ability to recognize unseen categories at test time. To overcome this, open-vocabulary detection (OVDet) emerged, aiming to localize objects described by arbitrary textual queries. Early foundational works, such as ViLD [4] and OWL-ViT [23] demonstrated that large-scale vision-language models like CLIP [28] could be distilled into detection frameworks to enable zero-shot generalization. Subsequent research has focused on enhancing this transferability through improved distillation strategies and domain-adaptive designs [19, 35, 41]. Despite their success in recognizing novel classes, these models are stateless, producing independent predictions per query without the capacity to resolve linguistic ambiguity or incorporate iterative user feedback. Parallel to OVDet, visual grounding focuses on the more granular task of localizing specific image regions corresponding to complex natural language phrases. GLIP [15] pioneered the unification of detection and phrase grounding by reformulating object detection as a word-region alignment task. This direction was further advanced by transformer-based architectures like Grounding DINO [18], DetCLIPv2 [40], and CoDet [20]. Recent models continue to refine these approaches using query mechanisms [3, 9, 42], semantic graph constraints [17], and category-specific knowledge distillation [21], allowing detectors to better understand nuanced textual descriptions and object-to-object relationships. MMOVD [10] and OVMR [22] extended these ideas with multimodal fusion, improving text-image interactions. While these models excel at category-level recognition, they struggle with one-of-many scenarios where multiple candidate objects share near-identical semantic features. In these cases, models often assign nearly identical confidence scores to all matching instances, failing to detect the specific object intended by the user. This limitation is particularly evident in examples where a model must identify one among several similar objects (e.g., second-third columns of Figure 4). Such ambiguity highlights a shortcoming of current open-vocabulary and grounding approaches.

2.2 Interactive Modeling

The idea of learning from user interaction has long been explored in content-based retrieval [6, 24, 31]. In these systems, users provide relevance feedback by marking retrieved results as positive or negative, allowing the model to iteratively refine the query representation, often using the Rocchio algorithm [30]. These early methods

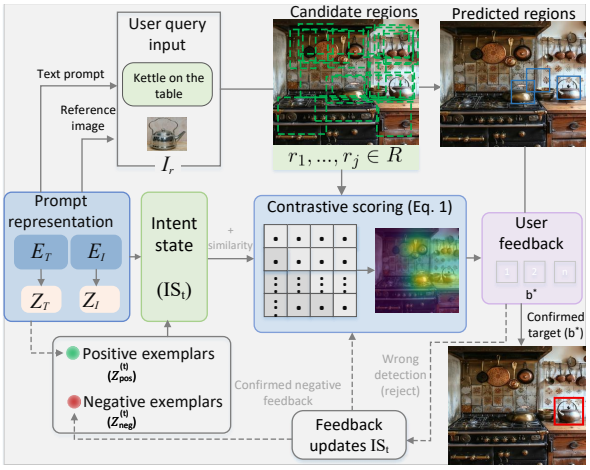


Figure 1: Overview of the proposed IntRec. It identifies a user-specified object through an interactive loop. An initial query is encoded to initialize the intent state, which containes both positive exemplars ($Z^{(t)}_{pos}$) and negative exemplars ($Z^{(t)}_{neg}$). This state guides the contrastive scoring module, which ranks all candidate regions in the target image based on their similarity to the positive exemplars and dissimilarity to the negative ones. The feedback updates the intent state, enabling the model to accurately localize the target object ($b^\star$).

Symbol	Definition
I_s, I_r	Target image used for retrieval, and an optional user-provided reference image.
p_0	Initial multi-modal prompt, consisting of a text description T_0 and/or a reference image I_r.
$b, b^\star$	Candidate bounding box and the final confirmed target object.
$R = \{r_1, \ldots, r_M\}$	Set of M candidate region feature vectors extracted from I_s.
E_T, E_I	Pre-trained text and image encoders (e.g., CLIP).
z_T, z_I, z_p	Embeddings of the text prompt, reference image, and fused query representation.
IS_t	Intent State at interaction turn t.
$Z^{(t)}_{pos}$	Positive exemplar within IS_t; stores embeddings representing user-confirmed cues.
$Z^{(t)}_{neg}$	Negative exemplar within IS_t; stores embeddings representing rejected.
$f_t = (b_j, s_t)$	Feedback tuple, where b_j is the selected region and $s_t \in \{\text{positive}, \text{negative}\}$ is the feedback type.

Table 1: Notation summary for the proposed IntRec.

demonstrated the power of feedback for refining retrieval results but were not designed for object localization within a single image. More recent work has introduced interaction into visual models in different ways. Diffusion-TTA [25] adapts vision encoders using test-time feedback, while interactive segmentation methods [34] show how user corrections can refine object boundaries. Closest to our work, Qin et al. [27] proposed an interactive learning framework for text-to-image person re-identification. However, while

effective for identity-matching across image galleries, such methods are not optimized for resolving the semantic ambiguity of general object categories in cluttered, open-world scenes. A common limitation persists across these domains: predictions are generally made independently for each query, with no mechanism to resolve distractors through a stateful dialogue. Our work addresses this gap by proposing IntRec that combines vision-language models with a contrastive feedback mechanism for user-guided object localization.

3 PROPOSED MODEL

Given a target image I_s, our goal is to localize a specific object $b^\star$ through a sequence of user interactions $t = 0, 1, \ldots, K$, rather than relying on predefined categories. At the initial turn $t = 0$, the user provides a query prompt p_0, which may include a text description T_0, a reference image I_r, or both. The model extracts a set of M candidate regions from I_s, represented by their feature embeddings $R = \{r_1, \ldots, r_M\}$, where each $r_i \in R^d$. At each subsequent turn $t > 0$, the user provides feedback $f_t = (b_j, s_t)$, where b_j is a candidate region and $s_t \in \{\text{positive}, \text{negative}\}$ indicates whether that region matches the intended target. Positive feedback confirms a user's target object, while negative specifies visual attributes to avoid. Ambiguity arises when multiple regions satisfy the initial query equally well, making user feedback essential for disambiguation.

3.1 Intent State (IS) Representation

In most retrieval models, user intent is represented as a single embedding vector obtained by encoding the input query. While simple, this formulation compresses all semantic cues into a single point and lacks a mechanism to learn from user rejections [37], limiting its ability to distinguish between similar candidates. We propose an Intent State (IS), a memory representation that evolves throughout the interaction. At any turn t, the IS is defined as a tuple of two exemplar sets: $IS_t = Z_{pos}^{(t)}, Z_{neg}^{(t)}$, where the first set stores positive anchors (embeddings representing user-confirmed cues or desired attributes), and the second stores negative constraints (embeddings of rejected attributes). The process begins by converting the initial prompt p_0 into the state IS_0. The text prompt T_0 and reference image I_r are encoded using CLIP encoders as $z_T = E_T(T_0)$ and $z_I = E_I(I_r)$, respectively. These are fused into a single query vector $z_p^{(0)} = \alpha \cdot z_T + (1 - \alpha) \cdot z_I$, where α is a balancing hyperparameter. This vector forms the first entry in the positive exemplar set, initializing the state as $Z_{pos}^{(0)} = \{z_p^{(0)}\}$, and $Z_{neg}^{(0)} = \emptyset$. By modeling intent as a distribution of exemplars rather than a single embedding, this representation enables more precise and stable retrieval.

3.2 Candidate Ranking Function

At each turn t, the model ranks all candidate regions R according to the current state IS_t. To achieve this, we propose a contrastive alignment score that acts as a exemplar-based classifier, measuring each candidate's relationship to both positive and negative exemplars. Formally, the score for a candidate region r_j is defined as

$$(r_j \mid IS_t) = \underbrace{\max_{z^+ \in Z_{pos}^{(t)}} \cos(r_j, z^+)}_{\text{Similarity to positive intent}} \underbrace{- \lambda \cdot \max_{z^- \in Z_{neg}^{(t)}} \cos(r_j, z^-)}_{\text{Penalty for negative intent}} \quad (1)$$

where λ controls the influence of negative constraints. The first term promotes regions that align closely with any positive exemplar. For instance, if z^+ contains embeddings for both a text prompt and a reference image, the model identifies candidate regions that match either cue. The second term penalizes regions similar to rejected exemplars (z^-), effectively creating low-scoring valleys around non-target concepts in the embedding space. Together, these terms enable fine-grained discrimination between visually similar objects.

3.3 Interactive State Update

The interactive state update is the core learning mechanism of our framework, allowing the model to learn from corrections and progressively converge on the intended target. Let r_j be the embedding of the region selected by the user at turn $t + 1$ with feedback $f_{t+1} = (b_j, s_{t+1})$. The update rule depends on the feedback type. (i) Negative feedback (s_{t+1} = negative): the rejected region's feature vector is added to the negative set ($r_j \rightarrow Z_{neg}$)

$$Z_{neg}^{(t+1)} = Z_{neg}^{(t)} \cup \{r_j\} \quad (2)$$

while the positive set remains unchanged $Z_{pos}^{(t+1)} = Z_{pos}^{(t)}$. This update teaches the model which visual features to avoid (e.g., distractors). (ii) Positive feedback (s_{t+1} = positive): if the user confirms a region (object) or provides a new textual prompt (encoded as z_{new}), the corresponding feature is added to the positive exemplar

$$\begin{aligned} Z_{pos}^{(t+1)} &= Z_{pos}^{(t)} \cup \{r_j, z_{new}\} \\ Z_{neg}^{(t+1)} &= Z_{neg}^{(t)} \quad \text{(Unchanged)} \end{aligned} \quad (3)$$

the feedback feature can be an image region r_j, or the encoded new text prompt. The updated $IS_{t+1} = (Z_{pos}^{(t+1)}, Z_{neg}^{(t+1)})$ is then used in the next turn to re-score the candidates, using Eq. 1. This loop continues until the target is confirmed. Through this process, the model learns not only which object is wrong, but also which visual features are misleading, an ability absent from most open-world detectors. In summary, IntRec begins with a query encoded into text and image embeddings, forming the initial intent state (IS_0). The model generates candidate regions for the target image and evaluates them using the contrastive scoring function (Eq. 1). The top-ranked region is displayed to the user, who provides feedback to confirm or reject the result. This feedback updates the intent state ($\text{IS}_t \rightarrow \text{IS}_{t+1}$), progressively refining subsequent predictions. The full procedure is summarized in Algorithm 1, Table 1.

3.4 Theoretical analysis of ambiguity resolution

Most open-vocabulary detectors (e.g., OWL-ViT, Grounding DINO), simply returns the region (object of interest, $r^\star$) with the highest similarity to the query T, and defined by

$$b^\star = f_{\text{standard}}(T, R) = \underset{b_j}{\text{argmax}}(\text{sim}(E_T(T), r_j)) \quad (4)$$

An ambiguity condition exists if there is at least one distractor object $r_d \neq r^\star$, such that its similarity to the query is greater than or equal to the true target's similarity $\text{sim}(E_T(T), r_d) \geq \text{sim}(E_T(T), r^\star)$. Under this condition, the stateless model will incorrectly select r_d and has no mechanism for correction. In contrast, our framework is designed to resolve such ambiguity. At the initial turn ($t = 0$), our score function behaves identically to the standard case

Algorithm 1 Our proposed dynamic intent search algorithm

Require: Target image I_s, prompt $p_0 = \{T_0, I_r\}$, max interaction turns K.
Ensure: Localized target $b^\star$.

Extract candidate region embeddings $R = \{r_1, \ldots, r_M\}$ from I_s
Encode query $z_T = E_T(T_0)$, $z_I = E_I(I_r)$, $z_p^{(0)} = \alpha z_T + (1-\alpha) z_I$
Initialize intent state: $Z_{pos}^{(0)} = \{z_p^{(0)}\}$, $Z_{neg}^{(0)} = \emptyset$
for $t = 0, 1, \ldots, K-1$ **do**
  for $r_j \in R$ **do**
    $S_{pos} = \max_{z^+ \in Z_{pos}^{(t)}} \cos(r_j, z^+)$
    $S_{neg} = \max_{z^- \in Z_{neg}^{(t)}} \cos(r_j, z^-)$ **if** $Z_{neg}^{(t)} \neq \emptyset$
    $S(r_j) = S_{pos} - \lambda S_{neg}$ ▷ Contrastive ranking (Eq. 1)
  end for
  Present top-k candidates to the user and receive feedback $f_{t+1} = (b_j, s_{t+1})$
  if s_{t+1} = positive-confirmation **then**
    return $b^\star = b_j$
  else if s_{t+1} = positive-refinement **then**
    $Z_{pos}^{(t+1)} = Z_{pos}^{(t)} \cup \{r_j \text{ or } z_{\text{new}}\}$
  else if s_{t+1} = negative **then**
    $Z_{neg}^{(t+1)} = Z_{neg}^{(t)} \cup \{r_j\}$
  end if
end for

$S(r_j \mid IS_0) = \text{sim}(r_j, E_T(T))$, and may also select the distractor r_d. After receiving negative feedback on r_d, the intent state updates to $IS_1 = (Z_{pos}^{(1)}, Z_{neg}^{(1)}) = (\{E_T(T)\}, \{r_d\})$. The new scores for the distractor and the target become $S(r_d|IS_1) = \text{sim}(r_d, E_T(T)) - \lambda$, $S(r^\star|IS_1) = \text{sim}(r^\star, E_T(T)) - \lambda \text{sim}(r^\star, r_d)$ The ambiguity is resolved if $S(r^\star|IS_1) > S(r_d|IS_1)$, which holds

$$\lambda(1 - \text{sim}(r^\star, r_d)) > \text{sim}(r_d, E_T(T)) - \text{sim}(r^\star, E_T(T))$$

As the right-hand side is a small non-negative, and $1 - \text{sim}(r^\star, r_d)$ is a positive constant (since $r^\star \neq r_d$), the inequality can always be satisfied with a suitable penalty weight λ. This proves that our contrastive mechanism is guaranteed to resolve the ambiguity. Consider a query for "the most front cup" in a scene with multiple similar cups. A stateless model may correctly identify a cup in the front row but select the wrong one, as several candidates have nearly identical high scores. However, our interactive model, after detecting the wrong object, follow-up by "not this, the cup near the left bottle". The model then, parses "not this" as a negative constraint. Using this feedback, we update both sets of the intent state simultaneously. The visual features of the rejected cup are added to the negative constraint set (Z_{neg}), while the new clue, is encoded and added to the positive anchor set (Z_{pos}). In the subsequent search, the model re-ranks all candidates based on this richer, more constrained intent state. This corrective feedback and update both positive and negative constraints is what separates our model from current open-vocabulary detectors.

4 EXPERIMENTS

4.1 Architecture and Implementation Details

We use the pre-trained, frozen CLIP ViT-B/16 encoders [28] for both text (E_T) and image (E_I) feature extraction. All textual prompts and image region features are projected into a L2-normalized embedding space of dimension d=512. To generate object proposals (candidate regions), we follow recent works [10, 22, 46] and employ CenterNet2 [47] with a ResNet-50 backbone pre-trained on the ImageNet-21k dataset [29]. The weights of both the CLIP and CenterNet2 models are kept frozen during inference. For any target image I_s, we use the detector's class-agnostic proposal to generate $M = 100$ candidate bounding boxes. The features for these proposals, $R = \{r_1, \ldots, r_{100}\}$, are then extracted using frozen CLIP image encoder. To measure how the interaction process improves detection performance, we adopt the following evaluation protocol. The interaction is simulated for a maximum of $K = 2$ turns. Turn-0 (baseline): for each query, we first evaluate the model's initial, non-interactive prediction, and compute the AP @ Turn-0 score. Turn-1 (post-feedback): if the model's top-1 prediction at Turn-0 was incorrect, we provide negative feedback on that specific object. The model updates its intent state and generates a refined ranked list of objects. We then compute the AP @ Turn-1. The hyperparameters are set as: α (modality fusion) = 0.6, λ (negative penalty weight) = 1.0, and M (number of candidate regions) = 100.

4.2 Dataset and Evaluation metrics

We evaluate our framework on two large-scale object detection datasets: LVIS v1 [5] and Objects365 [33]. Following the ViLD protocol [4], we treat the 866 common and frequent LVIS categories as base classes, and the 337 rare categories as novel classes to assess open-vocabulary generalization. Objects365, which contains 365 categories across more than 600k images, is used for transfer detection experiments. We report mean Average Precision (AP) across all categories, as well as AP(r), AP(c), and AP(f) for rare, common, and frequent subsets, respectively. Our goal is to improve AP(r), without significantly reducing overall AP. To specifically evaluate our model's ability to handle ambiguous retrieval scenarios, we constructed a challenging subset from the LVIS, which we call LVIS-Ambiguous. This benchmark is designed to include cases where multiple visually similar objects of the same category appear in the same image, causing standard open-vocabulary detectors to make incorrect predictions. We used a pre-trained Grounding DINO [18] as a probe. For each ground-truth object (in an image I) with bounding box b_{gt} and category label c_{gt}, we generate a simple category-level text prompt (i.e, T=a photo of a c_{gt}). This triplet (I, T, b_{gt}) forms one candidate sample for the benchmark. We then input I and T into the frozen Grounding DINO, which outputs a ranked list of N detected objects with bounding boxes and confidence scores, $D = [(b_1, s_1), (b_2, s_2), \ldots, (b_N, s_N)]$. Next, we compute the IoU between the ground-truth box b_{gt} and every predicted box $b_i \in D$, and identify $b_{gt-pred}$ with the highest IoU to b_{gt} (let its rank in the list be k). We then inspect all predictions ranked higher than this true target (i.e., b_j with rank $j < k$). A prediction b_j is a distractor if it satisfies both conditions: it has a low overlap with the true object (i.e., $IoU(b_j, b_{gt}) < 0.5$), and, it belongs to the same category c_{gt} as the true object (verified via ground-truth annotations). If at least one such distractor is ranked higher than the true target, the sample (I, T, b_{gt}) is an ambiguous case, and added to the LVIS-Ambiguous benchmark. This procedure yields a curated subset of fine-grained, visually confusing cases.

For this experiment we report two scores for our model. In AP @ Turn-0, we first run our model in a non-interactive mode, using only

Method	backbone	AP (r / c / f)
Detic [46]	ResNet-50	27.4 (18.6 / - / -)
DetPro [2]	ResNet-50	26.8 (20.3 / 26.5 / 28.9)
RegionCLIP [45]	ResNet-50	28.2 (17.1 / 27.4 / 34.0)
ViLD [4]	ResNet-50	25.8 (16.6 / 24.6 / 30.1)
CCKT-Det [43]	ResNet-50	27.1 (18.2 / - / -)
F-VLM [12]	ResNet-50	24.2 (18.6 / - / -)
BARON [38]	ResNet-50	29.5 (23.2 / 29.6 / 33.8)
VLDet [16]	ResNet-50	30.1 (21.7 / 29.8 / 34.3)
DVDet [9]	ResNet-50	31.2 (23.1 / 31.2 / 35.4)
NRAA [26]	ResNet-50	27.8 (21.2 / 27.0 / 31.5)
MMOVD [10]	ResNet-50	31.3 (19.8 / 31.5 / 34.2)
CoDet [20]	ResNet-50	31.7 (24.5 / 31.0 / 35.4)
LBP [13]	ResNet-50	29.9 (24.7 / 29.5 / 32.8)
OVMR [22]	ResNet-50	33.1 (23.5 / 32.1 / 36.9)
CAKE [21]	ResNet-50	34.9 (25.0 / 34.8 / 38.4)
MIC [36]	ResNet-50	33.8 (22.1 / 33.9 / 40.0)
IntRec- T	ResNet-50	**35.0 (24.7** / 34.1 / 38.0)
IntRec- MM	ResNet-50	**35.4 (25.6** / 34.7 / 38.2)

Table 2: Comparison results on LVIS using ResNet-50 detector backbone. AP(r) and AP are the main evaluation metric for this experiment. Rows for our models are highlighted, representing results from text, vision, and multimodal.

Method	object365			COCO		
	AP	AP-50	AP(r)	AP	AP-50	AP-75
DetPro [2]	11.7	18.2	9.8	34.9	53.8	37.4
BARON [38]	12.7	20.3	10.2	36.2	55.7	39.1
CoDet [20]	13.8	20.6	10.5	37.4	55.1	39.5
CCKT-Det [43]	13.4	19.7	11.5	38.5	53.2	42.1
OVMR [22]	12.3	19.5	10.7	37.6	54.7	40.3
DetLH [8]	14.6	21.2	11.9	38.9	56.5	41.6
IntRec (Turn-0)	13.8	21.4	11.5	38.6	56.0	41.3
IntRec (Turn-1)	**17.2**	**23.9**	**14.7**	**41.5**	**57.8**	**43.7**

Table 3: Transfer detection result. All models are trained on the LVIS/ImageNet-21K, and evaluated on Object365 and COCO. Our model is shown at both Turn-0 and Turn-1.

Model	AP on LVIS-Ambiguous
CoDet [20]	13.9
OVMR [22]	14.5
IntRec (Turn-0)	14.8
IntRec (Turn-1)	**22.7**

Table 4: Performance on the LVIS-Ambiguous Benchmark. While existing state-of-the-art models perform well on LVIS overall, their performance collapses under ambiguity. In this challenging scenario, our interactive model demonstrates a substantial recovery and clear performance advantage.

the initial prompt (without using feedback), then calculate the AP score. In AP @ Turn-1, for each sample, we take the top-1 prediction from the Turn-0 and provide it as a negative feedback signal. Our model updates its intent state and produces a new, refined ranked list. We calculate the AP score on this new list.

4.3 Comparison with State-of-the-Art Methods

4.3.1 Open world object detection. We evaluate our model on the LVIS and compare it against recent state-of-the-art open-vocabulary detectors. To ensure a fair comparison, all models use a ResNet-50 [7] backbone. Performance is reported in Table 2 using mean Average Precision (AP), with a focus on rare classes AP_r. While methods such as MMOVD [10], MIC [36], and OVMR [22] mitigate ambiguity by using multiple (e.g., 10) text and image prompts, our model takes a simpler strategy. Given a single query, our model performs an initial detection pass to identify the top-1 ranked object, which serves as a distractor. It then performs a second inference pass, where all candidate regions are re-ranked using our contrastive scoring function, with the original query as a positive anchor and the top-1 prediction from the first pass as a negative constraint. The final ranked list from this second pass is used to compute our model's AP scores. As shown in Table 2, our text-only model sets a new state-of-the-art on LVIS. When augmented with a reference image, performance improves further, surpassing the previous best methods on AP_r, including CAKE (25.0), CoDet (24.5), LBP (24.7). Although some recent models, such as MM-GDINO [44], Grounding DINO [18], LLMDet [3], DesCo [14], and DITO [11], using larger backbones (e.g., Swin-T and ViT-S/16), their AP(r) remains limited (23.6 / 10.1 / 26.0 / 19.6 / 26.2, respectively), and they require significantly more computational resources.

4.3.2 Analysis on the LVIS-Ambiguous benchmark. Our model is designed to recover from initial mispredictions through interaction. Consider an image densely packed with several small ceramic cups, where the target is "the cup at the very front". Most detectors, which relies on a single similarity score, often fail in this cases [17, 35, 39]. Because all cups appear visually similar, their similarity scores are nearly identical, making the top-ranked prediction unstable and incorrect (see Section 3.4). With a simple feedback such as "not this one, the cup near the left bottle", our model updates its intent state to suppress the rejected region and refocus on the correct target. To evaluate this behavior, we use our LVIS-Ambiguous benchmark (see Section 4.2). We compare the performance of top-performing baselines against our model in two modes: Turn-0, representing the initial non-interactive prediction, and Turn-1, representing the prediction after one round of feedback. The results in Table 4, reveal that all models perform poorly at Turn-0, confirming the difficulty of this benchmark. However, after a single corrective interaction, our model's performance at Turn-1 rises to 22.7, an improvement of +7.9, demonstrating its ability to recover from initial mispredictions and handle visual ambiguity.

4.3.3 Transfer detection setting. To evaluate our model generalization capabilities, we test it in a zero-shot transfer detection setting. The model, which has been trained on LVIS/ImageNet-21k, is evaluated directly on the Objects365 and COCO datasets without any

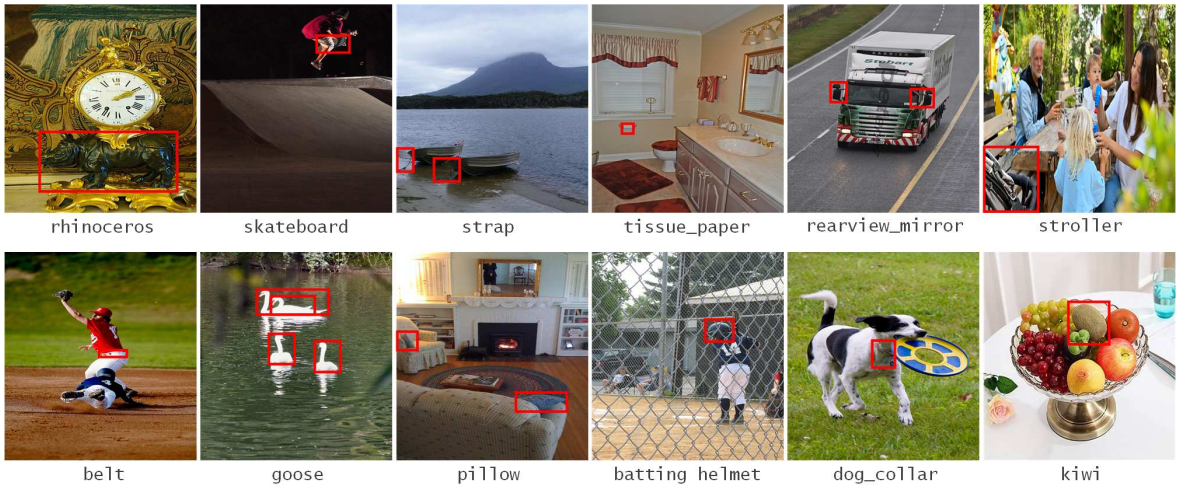


Figure 2: Qualitative examples of our proposed model detecting rare categories in the LVIS validation set using textual prompt.

fine-tuning. AP(r) is evaluated on least frequent 25% categories in the Objects365 training set. We evaluate our model in two modes: Turn-0 performance, where the prediction is based on the initial text prompt., and Turn-1 which is the result after applying our correction (feedback) mechanism, where the top-1 prediction from Turn-0 is used as a negative constraint to re-rank all candidates. The results are shown in Table 3. Our model at Turn-0 performs comparably with other baselines, but, after a single correction pass, our Turn-1 performance shows a significant boost across all metrics on both datasets, specially on the rare categories.

4.3.4 Alignment strategy: Local vs. Global ranking. One of the key mechanisms in our framework is the ranking function (Eq. 1), which evaluates each object (candidate region r_j) independently using a local contrastive score. To justify whether this local strategy is indeed optimal, we compare it against a global assignment baseline based on optimal transport. For this baseline, we replace our scoring function at each interactive turn t with Sinkhorn-based algorithm [1], and construct a cost matrix between the positive anchors Z_{pos} and all candidate regions R, where the cost is inversely proportional to their similarity. Negative constraints, Z_{neg}, are used as a high cost (penalty) for undesirable regions in this matrix. The Sinkhorn algorithm then computes a soft assignment plan, from which the final ranking of candidates is derived. We evaluate both variants on the LVIS-Ambiguous benchmarks over a maximum of $K = 2$ interactive turns, focusing on novel/rare category performance. As shown in Figure 5, our model consistently achieves a higher final AP, surpassing the Sinkhorn-based global assignment variant by +1.4 (at turn=2). We hypothesize that this advantage arises because local scoring can better exploit the sharp, localized signal from a negative constraint, whereas global assignment tends to dilute this signal across the entire candidate set.

4.3.5 Qualitative results. Figure 3 shows visual attention maps generated by our model for several complex queries, alongside results from MIC [36] and OVMR [22]. These visualizations illustrate how well each model captures fine-grained linguistic attributes in the

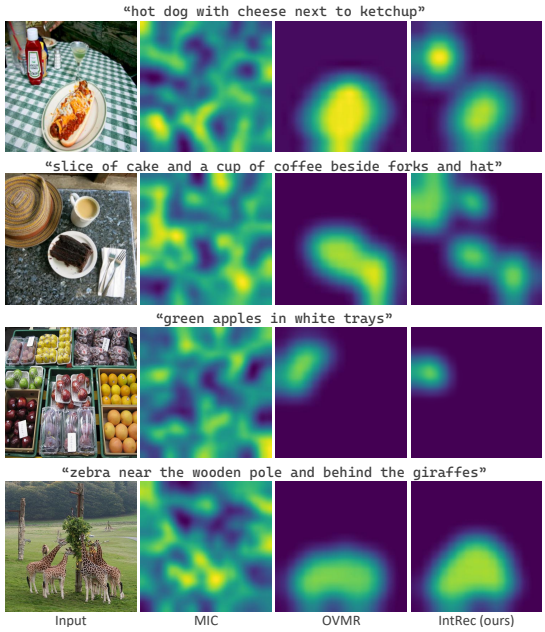


Figure 3: Localization comparison. While baseline models produce diffuse or imprecise heatmaps, our model generates sharp, accurate localizations that correctly ground all semantic components of the prompt. For instance, it correctly distinguishes the green apples from other fruit and localizes the zebra despite the presence of nearby giraffes.

image. Our model produces sharper and more focused attention regions, indicating a stronger correspondence between the textual query and the visual evidence. Specially, the MIC model generates diffuse and scattered attention, indicating difficulty in grounding the full query. While OVMR can identify the correct object region, still displays coarse and imprecise focus. For example, in the query "green apples in white trays", both baselines fail to distinguish

Figure 4: Detection comparison in cluttered scenes. Blue boxes indicate detections of novel/rare categories, red boxes indicate detections of known (base) categories. While the baseline models frequently produce a redundant and overlapping detections for novel objects, our model generates accurate predictions, suppress duplicate detections in dense environments.

between visually similar objects, whereas our model provides a focused heatmap that aligns with the target objects described in the text (detection result for our model is given in Figure 2).

To further illustrate this behavior, Figure 4 compares the final bounding box predictions of our model against several baselines. In these examples, red boxes indicate detections of base (known) categories, and blue boxes denote detections of novel/rare categories. Although the baselines are capable of detecting novel objects, they often produce redundant and overlapping boxes (i.e., multiple imprecise detections corresponding to the same instance). This is a failure mode for open-vocabulary detectors in dense scenes [3, 18]. In contrast, our model yields more accurate detections, producing well-localized boxes that correspond to the intended objects.

4.4 Ablation Studies

We conduct a series of ablation studies to validate the key components of our framework: the intent state and our contrastive ranking function. For this, we evaluate several variants of our model on the LVIS benchmark, reporting the AP score achieved after a maximum of $K = 2$ interactive turns in Figure 6. We compare the performance of our model with its variants. i) Our proposed model, which uses the intent state and the full contrastive score, $\max(pos) - \lambda \cdot \max(neg)$. ii) w/o negative feedback, where we set $\lambda = 0$, then the score becomes a similarity search to the positive anchors $\max(pos)$. iii) w/ averaging score, where we replace our max

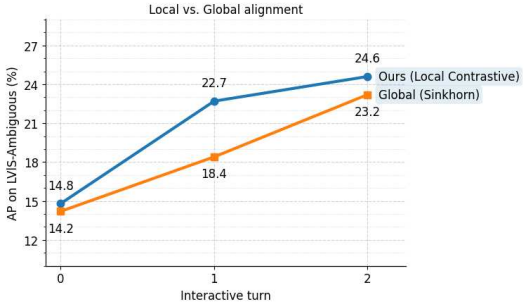


Figure 5: Analysis of Local vs. Global alignment on the LVIS-Ambiguous benchmark over two interactive turns. Both models start with a low AP at Turn 0 (initial prompt). However, after the first corrective feedback (Turn 1), our model increases by +7.9, outpacing the +4.3 gain of the global baseline.

operators with an averaging of exemplars, then the score becomes $cos(r_j, \text{avg}(Z_{pos})) - \lambda \cdot cos(r_j, \text{avg}(Z_{neg}))$. iv) w/o intent state (stateless), that at each turn t, the intent state is reset and only contains the initial prompt and the most recent user feedback. The results show that our full model significantly outperforms all versions. The most performance degradation occurs in the stateless variant (a drop of -10.8 AP), proving that the intent state is the most critical

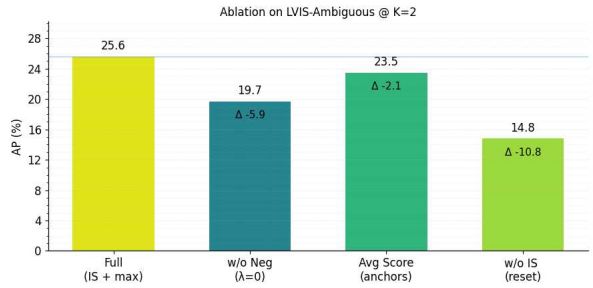


Figure 6: Ablation study of our model components on the LVIS benchmark. The results show the final AP after $K = 2$ interactive turns. Removing either the intent state (IS) memory or the negative feedback mechanism causes a dramatic drop in performance, demonstrating that both are essential for effective ambiguity resolution. Our full model significantly outperforms all degraded variants.

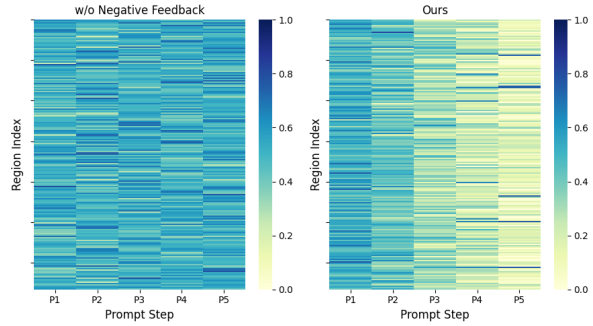


Figure 7: Visualization of the contrastive refinement process. This shows the normalized scores of 100 candidate regions (y-axis) over 5 interactive steps (x-axis) for an ambiguous query with multiple distractors. (Left) Without our contrastive mechanism ($\lambda = 0$), the model is confused; scores for many distractor regions stay high throughout the interaction. (Right) Our full model uses negative feedback to suppress the scores of irrelevant regions (light yellow) while specifying the score of the true target (dark blue).

component for resolving complex ambiguity. Disabling negative feedback also weaken the model's performance (a drop of -5.9 AP), confirming that our contrastive learning from rejection is essential for disambiguation. Finally, our max-operator for ranking function outperforms the averaging strategy. Additionally, we compare our full model against a key ablation, which is disabling the negative feedback, and the result is presented in Figure 7. For this experiment, we select a an example image from LVIS-Ambiguous benchmark with many similar objects. Then start with an initial text prompt (P1); for steps P2 through P5, simulate a user providing negative feedback on the current highest-scoring incorrect object. At each of the 5 steps, record the scores for all 100 candidate regions for both our model and its baseline (w/o negative feedback). As can be seen, our baseline model identifies many plausible regions at the start (P1), but because it cannot learn from negative feedback, the scores for the wrong objects (distractors) remain high throughout the interaction (P2-P5). However, our full model, at the start (P1), it is also confused and gives high scores to many regions. But as the user provides negative feedback in subsequent steps (P2, P3, P4), we can see the model learning, and the scores for the incorrect regions are suppressed and the correct target is identified.

m	AP (r/c/f)
25	34.5 (25.0 / 34.2 / 37.8)
50	34.9 (25.7 / 34.3 / 38.0)
100	35.4 (25.6 / 34.7 / 38.2)
200	34.9 (24.3 / 34.0 / 38.1)

(a) Number of region candidate (optimal: $m = 100$)

λ	AP (r/c/f)
0.1	33.7 (24.2 / 33.6 / 37.5)
0.5	35.6 (25.0 / 34.5 / 37.9)
1.0	35.4 (25.6 / 34.7 / 38.2)
1.5	35.1 (24.9 / 34.2 / 38.0)

(b) Contrastive loss weight (optimal: $\lambda = 1$)

Table 5: Hyperparameter sensitivity on LVIS/ResNet-50.

4.4.1 Hyperparameter sensitivity. Table 5 analyzes the effect of two key hyperparameters: the number of region candidates, m and the weight of the contrastive loss, λ. When $m = 25$, the model achieves the lowest AP on rare categories (25.0), due to insufficient proposal coverage. Increasing to $m = 100$ provides the best result, however, setting $m = 200$ degrades performance on rare categories by 1.3%, likely due to the inclusion of low-quality or noisy regions. For contrastive loss, $\lambda = 0.5$ improves overall AP (35.6), but have slightly weaker performance on rare categories. We find that $\lambda \in [0.5, 1.0]$ provides balanced results, we set it to 1. In terms of efficiency, a single interactive turn introduces only ≈29 ms of additional computation on an NVIDIA RTX 3090 GPU, which corresponds to less than 15% of the total inference time. This confirms that the proposed feedback mechanism offers substantial accuracy gains with minimal computational overhead.

5 CONCLUSION AND FUTURE WORK

In this work, a new framework is proposed for interactive object localization that can overcome a key limitation of current open-vocabulary detectors, which often fail on ambiguous queries involving multiple similar objects. We addressed this by reframing object retrieval as a stateful learning process, allowing the model to iteratively refine the user intent. The effectiveness of our model is driven by its intent state module, which accumulates positive and negative feedback, and a contrastive alignment function that leverages this memory to disambiguate between visually similar objects. Our experiments demonstrate that even a single corrective feedback significantly improves retrieval accuracy on standard open-world benchmarks, and especially on the challenging LVIS-Ambiguous subset. However, our model remains limited by its reliance on the initial set of candidate regions. If the detector fails to generate a bounding box for the true target (e.g., when the object is too small or heavily occluded), the interactive refinement process cannot recover it. In future work, we plan to explore mechanisms for updating or refining the candidate proposals based on user feedback.

REFERENCES

[1] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).

[2] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14084–14093.

[3] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. 2025. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14987–14997.

[4] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. (2022).

[5] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5356–5364.

[6] Donna Harman. 1992. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. 1–10.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Jiaxing Huang, Jingyi Zhang, Kai Jiang, and Shijian Lu. 2024. Open-vocabulary object detection via language hierarchy. *arXiv preprint arXiv:2410.20371* (2024).

[9] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. 2024. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *International Conference on Learning Representations* (2024).

[10] Prannay Kaul, Weidi Xie, Andrew Zisserman, and . 2023. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*. 15946–15969.

[11] Dahun Kim, Anelia Angelova, and Weicheng Kuo. 2024. Region-centric Image-Language Pretraining for Open-Vocabulary Detection. In *European Conference on Computer Vision*. 162–179.

[12] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. 2023. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *International Conference on Learning Representations* (2023).

[13] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. 2024. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16678–16687.

[14] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. 2023. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems* 36 (2023), 37511–37526.

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

[16] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. 2023. Learning object-language alignments for open-vocabulary object detection. *International Conference on Learning Representations* (2023).

[17] Xin Lin, Chong Shi, Zuopeng Yang, Haojin Tang, and Zhili Zhou. 2025. SGC-Net: Stratified Granular Comparison Network for Open-Vocabulary HOI Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 4539–4549.

[18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*. 38–55.

[19] Yang Liu, Feng Hou, Yunjie Peng, Gangjian Zhang, Yao Zhang, Dong Xie, Peng Wang, Yang Zhang, Jiang Tian, Zhongchao Shi, et al. 2025. DoGA: Enhancing Grounded Object Detection via Grouped Pre-Training with Attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5658–5666.

[20] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. 2024. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems* (2024).

[21] Shiyuan Ma, Donglin Qian, Kai Ye, and Shengchuan Zhang. 2025. Cake: Category aware knowledge extraction for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5982–5990.

[22] Zehong Ma, Shiliang Zhang, Longhui Wei, and Qi Tian. 2024. OVMR: Open-Vocabulary Recognition with Multi-Modal References. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16571–16581.

[23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*. 728–755.

[24] Ryoya Nara, Yu-Chieh Lin, Yuji Nozawa, Youyang Ng, Goh Itoh, Osamu Torii, and Yusuke Matsui. 2024. Revisiting relevance feedback for clip-based interactive image retrieval. In *European Conference on Computer Vision*. 1–16.

[25] Mihir Prabhudesai, Tsung-Wei Ke, Alex Li, Deepak Pathak, and Katerina Fragkiadaki. 2023. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback. *Advances in Neural Information Processing Systems* 36 (2023), 17567–17583.

[26] Sunyuan Qiang, Xianfei Li, Yanyan Liang, Wenlong Liao, Tao He, and Pai Peng. 2024. Open-Vocabulary Object Detection via Neighboring Region Attention Alignment. *arXiv preprint arXiv:2405.08593* (2024).

[27] Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. 2025. Human-centered Interactive Learning via MLLMs for Text-to-Image Person Re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14390–14399.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. 8748–8763.

[29] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972* (2021).

[30] Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971).

[31] Yong Rui, Thomas S Huang, and Shih-Fu Chang. 1999. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation* 10, 1 (1999), 39–62.

[32] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, Michael Felsberg, Dacheng Tao, and Xuelong Li. 2025. From Missing Pieces to Masterpieces: Image Completion with Context-Adaptive Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 7 (2025), 6073–6087.

[33] 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 *Proceedings of the IEEE/CVF international conference on computer vision*. 8430–8439.

[34] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. 2022. Reviving iterative training with mask guidance for interactive segmentation. In *IEEE international conference on image processing*. 3141–3145.

[35] Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. 2024. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844* (2024).

[36] Zhao Wang, Aoxue Li, Fengwei Zhou, Zhenguo Li, and Qi Dou. 2024. Open-Vocabulary Object Detection with Meta Prompt Representation and Instance Contrastive Optimization. *arXiv preprint arXiv:2403.09433* (2024).

[37] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2025. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038* (2025).

[38] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 2023. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15254–15264.

[39] Xing Xi, Yangyang Huang, Ronghua Luo, and Yu Qiu. 2025. OW-OVD: Unified Open World and Open Vocabulary Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 25454–25464.

[40] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. 2023. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23497–23506.

[41] Masoumeh Zareapoor, Pourya Shamsolmoali, and Yue Lu. 2024. Learning Region-Word Alignment with Attentive Masking for Open-Vocabulary Object Detection. In *NeurIPS 2024 Workshop on Open-World Agents*.

[42] Masoumeh Zareapoor, Pourya Shamsolmoali, and Yue Lu. 2025. BiMAC: Bidirectional Multimodal Alignment in Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 22290–22298.

[43] Chuhan Zhang, Chaoyang Zhu, Pingcheng Dong, Long Chen, and Dong Zhang. 2025. Cyclic Contrastive Knowledge Transfer for Open-Vocabulary Object Detection. *International Conference on Learning Representations* (2025).

[44] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. 2024. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361* (2024).

[45] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16793–16803.

[46] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*. 350–368.

[47] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2021. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461* (2021).