LIHE: LINGUISTIC INSTANCE-SPLIT HYPERBOLIC-EUCLIDEAN FRAMEWORK FOR GENERALIZED WEAKLY-SUPERVISED REFERRING EXPRESSION COMPREHENSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing Weakly-Supervised Referring Expression Comprehension (WREC) methods, while effective, are fundamentally limited by a one-to-one mapping assumption, hindering their ability to handle expressions corresponding to zero or multiple targets in realistic scenarios. To bridge this gap, we introduce the Weakly-Supervised Generalized Referring Expression Comprehension task (WGREC), a more practical paradigm that handles expressions with variable numbers of referents. However, extending WREC to WGREC presents two fundamental challenges: supervisory signal ambiguity, where weak image-level supervision is insufficient for training a model to infer the correct number and identity of referents, and semantic representation collapse, where standard Euclidean similarity forces hierarchically-related concepts into non-discriminative clusters, blurring categorical boundaries. To tackle these challenges, we propose a novel WGREC framework named Linguistic Instance-Split Hyperbolic-Euclidean (LIHE), which operates in two stages. The first stage, Referential Decoupling, predicts the number of target objects and decomposes the complex expression into simpler sub-expressions. The second stage, Referent Grounding, then localizes these sub-expressions using HEMix, our innovative hybrid similarity module that synergistically combines the precise alignment capabilities of Euclidean proximity with the hierarchical modeling strengths of hyperbolic geometry. This hybrid approach effectively prevents semantic collapse while preserving fine-grained distinctions between related concepts. Extensive experiments demonstrate LIHE establishes the first effective weakly supervised WGREC baseline on gRefCOCO and Ref-ZOM, while HEMix achieves consistent improvements on standard REC benchmarks, improving IoU@0.5 by up to 2.5%. The code is available at https://anonymous.4open.science/r/LIHE.

1 INTRODUCTION

Referring Expression Comprehension (REC) (Mao et al., 2016; Yu et al., 2016), also known as visual grounding, aims to localize objects in an image based on natural language expressions. REC has shown broad application potential in fields such as robotic navigation and image editing. However, existing REC methods heavily rely on instance-level annotations, which are expensive and labor-intensive to collect (Zhu et al., 2022; Deng et al., 2023). To alleviate this bottleneck, Weakly-Supervised REC (WREC) has emerged as a cost-effective alternative. WREC methods eliminate the need for bounding box supervision and instead learn to align vision and language features using only image–text pairs. Early approaches (Gupta et al., 2020; Liu et al., 2019c) typically employed two-stage pipelines, while recent single-stage frameworks (Jin et al., 2023; Luo et al., 2024; Cheng et al., 2025) have become dominant due to their superior efficiency, directly matching text with anchor features from pre-trained detectors through contrastive learning (Oord et al., 2018).

However, real-world scenarios are often more complex: a referring expression might correspond to multiple objects or no object at all, known as Generalized Referring Expression Comprehension (GREC) (He et al., 2023; Liu et al., 2023; Wu et al., 2024), which extends REC by allowing ex-

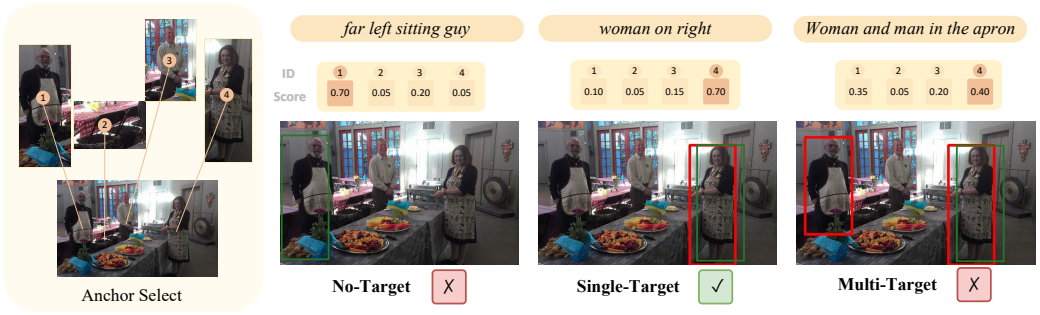


Figure 1: Limitations of current WREC methods. The ground truth is denoted by red bounding boxes, whereas green bounding boxes denote the predictions. Current WREC methods always select only the best anchor as output, failing to handle No-target and Multi-target cases (e.g., no red bounding box and two red bounding boxes).

pressions to describe no-target and multi-target cases (shown as Fig. 1). Accordingly, in this paper, we aim to solve a new task, Weakly-Supervised Generalized Referring Expression Comprehension (WGREC).

Extending WREC to the WGREC setting poses two fundamental challenges that need to be addressed to develop effective weakly supervised methods. The first is the ambiguity of the supervisory signal, where the winner-takes-all mechanism of previous WREC methods is structurally incapable of locating all relevant targets, erroneously returning only a single instance, as shown in Fig. 1. The second is the semantic representation collapse, which arises because conventional contrastive learning methods rely on Euclidean similarity. This assumes flat, one-to-one alignments, leading to suboptimal representations in multi-referent scenarios. For instance, an expression like "left person" may refer to both "left man" and "left woman" as shown in Fig. 3. Pulling both specific instances toward the same general anchor ("person") in Euclidean space unintentionally forces "man" and "woman" closer together, blurring their categorical boundaries.

To address these challenges, we propose a two-stage **L**inguistic **I**nstance-Split **H**yperbolic-**E**uclidean (LIHE) framework designed to effectively localize a variable number of objects from a single expression. First, in the **Referential Decoupling** stage, LIHE leverage a vision-language model (VLM) (Bai et al., 2023; Wang et al., 2024; Team, 2025) to infer the number of referring objects: if zero, it will skip the second stage; otherwise, it will decompose the expression into corresponding sub-expressions for each object. Second, the **Referent Grounding** stage employs a contrastive learning paradigm to train the model to localize each sub-expression. To prevent the semantic collapse inherent in this grounding process, LIHE integrates *HEMix*, a hybrid similarity module that replaces the standard Euclidean metric. HEMix synergizes the fine-grained alignment of Euclidean proximity with the structure-preserving properties of hyperbolic geometry (Ganea et al., 2018; Kim et al., 2023; Bdeir et al., 2024). It leverages the inherent capacity of Lorentzian hyperbolic space to model hierarchies by placing general concepts near the origin and specific instances outward, thereby preserving semantic distinctions. This design, which serves as a *plug-and-play* module, effectively prevents concept collapse and enhances generalization in complex referring scenarios with negligible computational overhead.

We conduct extensive experiments on the GREC datasets gRefCOCO (He et al., 2023) and Ref-ZOM (Hu et al., 2023), demonstrating that our method is the first weakly supervised framework to tackle the WGREC task effectively. Additionally, we validate the broader applicability of HEMix on REC benchmarks, including RefCOCO (Nagaraja et al., 2016), RefCOCO+ (Nagaraja et al., 2016), and RefCOCOg (Mao et al., 2016), and achieve the state-of-the-art performance. Our findings highlight the potential of combining REC with structured geometry to advance vision-language understanding.

2 Task Definition

The Weakly-Supervised Generalized Referring Expression Comprehension (WGREC) task aims to localize all image regions described by a given natural language expression using only weak supervision. Formally, given an image I and an expression T, the goal is to predict a set of bounding boxes, $\mathcal{B}^* = \{b_i \mid i \in k\}$, where each box b_i corresponds to a region in I that matches the expression T.

and the number of targets k is variable (zero, one, or more). Critically, no bounding box annotations are used during training. Each sample consists only of an image–text pair (I, T), without knowing which regions match the expression. Unlike WREC, which assumes every expression refers to a specific object, WGREC allows for expressions that have no matching region. To address this, we introduce a binary label $v \in \{0, 1\}$ for the training set, where $v = 0$ means no region in the image matches the expression, and $v = 1$ otherwise. This provides weak supervision to guide the model in learning to predict all matching regions. Our goal is to train a model that, using only these weak labels, can predict the complete set of referent boxes for inference.

3 METHOD

Extending WREC to WGREC is non-trivial and typically faces two major challenges: **(1) Cardinality Ambiguity**: WREC methods such as RefCLIP simplify the task by reducing it to an anchor–text matching problem. Specifically, they select the most relevant anchor from a predefined set $\mathcal{A}$ using:

$$a^* = \arg\max_{a \in \mathcal{A}} \phi(T, I, a), \tag{1}$$

where $\phi(T, I, a)$ denotes the similarity between the expression T, the image I, and anchor a. However, this max-selection strategy inherently assumes a single referent, making it unsuitable for WGREC, where the number of targets is unknown. **(2) Hierarchical Representation Collapse**: When a general expression T (e.g., "person") refers to multiple distinct sub-categories (e.g., a "man" and a "woman") as shown in Fig. 3, conventional contrastive learning in Euclidean space can conflate their representations. This blurs categorical boundaries and leads to a loss of semantic distinction.

To address these limitations, we propose **L**inguistic **I**nstance-Split **H**yperbolic-**E**uclidean (LIHE), a framework designed for WGREC. As shown in Fig. 2, LIHE consists of a *Referential Decoupling* stage, which decomposes complex expressions into single-instance queries, and a *Referent Grounding* stage, which detects all matching regions. In addition, we introduce a *HEMix* similarity to explicitly preserve hierarchical relationships.

3.1 REFERENTIAL DECOUPLING

In the context of WGREC, referring expressions often correspond to multiple visual entities. To address this, we reformulate the task by decomposing a multi-target expression into a set of single-target sub-expressions, allowing each referent to be localized independently and utilize the capabilities of VLM (Team, 2025) to judge whether the target exists. This strategy simplifies the multi-instance grounding problem and makes it more tractable under weak supervision.

We leverage the perceptual capabilities of large vision-language models (VLMs) (Team, 2025), which, although do not have grounding functions, exhibit strong visual understanding and language reasoning. Given an image I, a referring expression T, and a carefully designed prompt P, we input them into a VLM to obtain a set of simplified, instance-level expressions. To exploit the in-context learning capability of the VLM, we design a four-part prompt paired with the input image to guide the VLM in decomposing the original referring expression: (1) General instruction $\mathbf{P_G}$, describing the goal of splitting the expression into target-specific phrases; (2) Output constraints $\mathbf{P_C}$, specifying the format of each phrase; (3) In-context examples $\mathbf{P_E}$, providing annotated demonstrations to steer the VLM toward the desired behavior; (4) Input query $\mathbf{P_Q}$, which contains original referring expression together with explicit instructions; (5) Raw Image $\mathbf{I}$. Decomposition is formulated as

$$K, \mathcal{T}_\text{D} = \text{VLM}(\mathbf{P_G}, \mathbf{P_C}, \mathbf{P_E}, \mathbf{P_Q}, \mathbf{I}), \tag{2}$$

where K is the number of targets and $\mathcal{T}_\text{D} = \{t_1, t_2, \ldots, t_k\}$ is the set of sub-expressions generated by the model, each describing a distinct visual entity potentially present in the image. The prompts $\mathbf{P}$ guide the model to rewrite the original expression into concise, non-overlapping descriptions of individual targets, using brief instructions and a few examples. This decomposition mitigates issues such as irrelevance and ambiguity. Meanwhile, the prompt component $\mathbf{P_C}$ explicitly instructs the VLM to first output the number of target phrases K before listing them. This constraint helps mitigate common hallucination issues in VLMs, such as generating duplicate referring expressions for the same visual entity. When the VLM returns ($K = 0$), we interpret it as a no-target case, which naturally fits the open-ended setting of WGREC. More detailed prompt design, please see Appendix G.

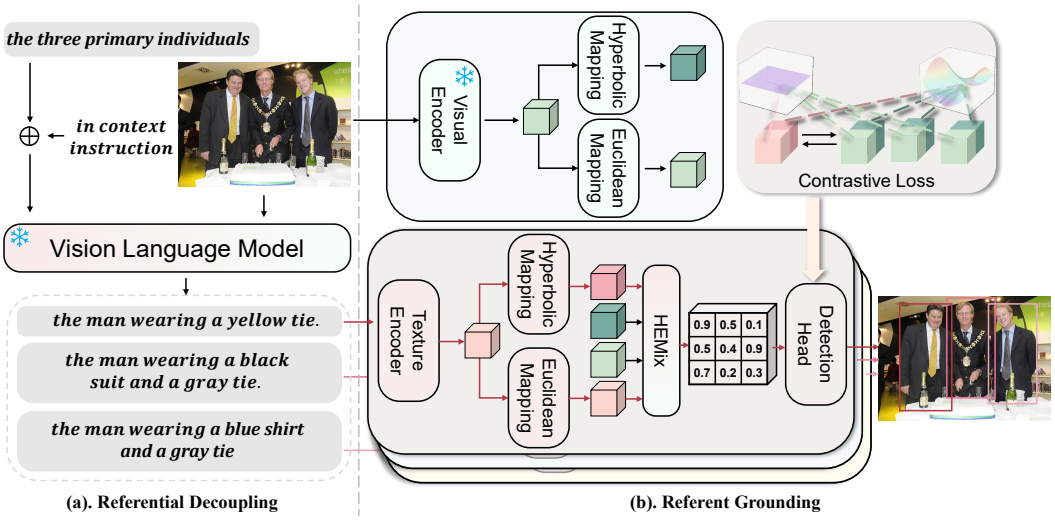


Figure 2: **The overall framework of LIHE**. (a). Referential Decoupling: VLM decomposed the referring expression into distinct short phrases for each target. (b). Referent Grounding: Each phrase is processed by a textual encoder, and the image by a visual encoder. Then the model filters anchors of low value and returns the best-matching one for bounding box prediction. The referent grounding stage is weakly supervised by the anchor-based contrastive loss.

3.2 Referent Grounding

After the decoupling stage, each decomposed referring expression t corresponds to a matched visual entity in the image I. Consequently, the task reduces to a conventional WREC problem, as defined in Eq. 1. Therefore, the referent grounding stage of our framework follows the structure and training strategy of previous WREC methods (Jin et al., 2023; Luo et al., 2024; Cheng et al., 2025), which have proven effective for WREC. Specifically, given an input image and a referring expression (the short phrases generated from the referential decoupling stage), the model uses a one-stage detector to extract visual feature maps, from which anchor features are obtained. We retain only the anchors from the last feature map layer—based on the assumption that most target objects in referring datasets are of moderate to large size and further filter them by confidence, typically keeping the top 10% of anchors. Each remaining anchor is projected into a joint semantic space, alongside the corresponding text embedding. The similarity between each anchor and the phrase is then computed using a hybrid similarity metric, which combines Euclidean and Hyperbolic similarity scores. Formally, we adopt the contrastive learning objective of RefCLIP (Jin et al., 2023), replacing its similarity function with our HEMix:

$$\mathcal{L}_c = -\log \frac{\exp\left(\text{HEMix}(f^i_{a_0}, f^i_t)/\tau\right)}{\sum_{n=0}^{N}\sum_{j=0}^{M} \mathbb{I}_{\neg(i=j \wedge n \neq 0)} \exp\left(\text{HEMix}(f^j_{a_n}, f^i_t)/\tau\right)} \tag{3}$$

where $f^i_{a_0}$ is the positive anchor for i-th image, f^i_t is the text embedding, τ is a temperature scalar, and HEMix denotes our proposed Euclidean-Hyperbolic hybrid similarity. This contrastive objective aligns the correct anchor-text pair while using both intra- and inter-image anchors as negatives. More details of HEMix formulation are provided in Sec. 3.3. Note that we filter the training dataset for the referent grounding stage by retaining only samples with validity flag $v = 1$ (i.e., at least one entity in I matches T).

3.3 HEMix

In previous WREC methods (Jin et al., 2023; Luo et al., 2024; Cheng et al., 2025), contrastive learning is typically driven by Euclidean similarity in a shared embedding space. However, this approach is limited in its ability to capture hierarchical semantics. For instance, as illustrated in Fig. 3, a phrase like *'left person'* may refer to both *'left man'* and *'left woman'* two visual entities in the image, which are semantically related but visually distinct. Euclidean similarity tends to cluster all such instances together, resulting in ambiguous localization. To better model semantic structure, we incorporate hyperbolic geometry using the Lorentz (hyperboloid) model. Hyperbolic

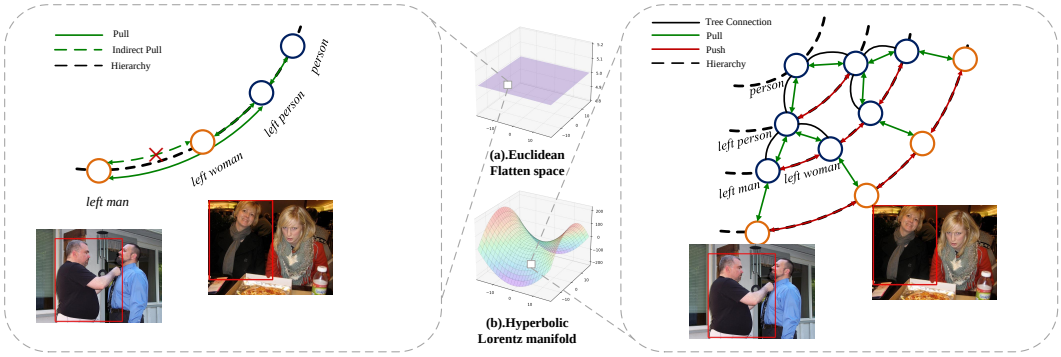


Figure 3: A simple illustration of (a) Euclidean flatten space and (b) hyperbolic Lorentz manifold in 3-dimensional space (Li et al., 2024). In Euclidean space, all nodes occupy a single, undifferentiated hierarchy, so parent and child entities share the same geometric scale. In contrast, the negative curvature of hyperbolic space naturally organizes nodes into concentric hierarchies: parent nodes reside closer to the manifold's apex, while their children are pushed farther outward, and different children at the same level are repelled from one another.

spaces naturally embed hierarchical relationships: pulling a parent concept (e.g., *'person'*) closer to the time axis increases angular separation among its children (e.g., *'man'* and *'woman'*), effectively preserving both generality and specificity. This aligns well with the hierarchical nature of referring expressions. More introduction and insights are in Appendix D.

For D-Dimension visual feature $f_v \in \mathbb{R}^D$ and text feature $f_t \in \mathbb{R}^D$, we use two types of similarity:

(1) **Euclidean similarity**: Same as refclip (Jin et al., 2023), the similarity is calculated by

$$\mathrm{Sim}_{\mathrm{E}}(f_v, f_t) = \langle f_v \mathbf{W}_{EV},\ f_t \mathbf{W}_{ET} \rangle, \tag{4}$$

where $\mathbf{W}_{EV}$ and $\mathbf{W}_{ET}$ are learnable linear mapping matrices, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product in Euclidean space.

(2) **Hyperbolic similarity (Lorentz model)**: We first map features into hyperbolic space as spatial components of $\tilde{f}_v, \tilde{f}_t$ in hyperbolic space:

$$\mathbf{z}_v = f_v \mathbf{W}_{HV} \in \mathbb{R}^D, \quad \mathbf{z}_t = f_t \mathbf{W}_{HT} \in \mathbb{R}^D, \tag{5}$$

where $\mathbf{W}_{HV}$ and $\mathbf{W}_{HT}$ are learnable linear projection matrices for hyperbolic space. Here, instead of using exponential maps to embed into hyperbolic space in the paper before, which is prone to unstable gradients, we adopt a learnable linear projection for both branches. This ensures smooth training and better compatibility with contrastive objectives. In a hyperbolic space of curvature κ, the calculation of hyperbolic similarity by the Lorentzian inner product is formulated as:

$$\mathrm{Sim}_{\mathrm{H}}(f_v, f_t) = \langle \tilde{f}_v, \tilde{f}_t \rangle_{\mathbb{H}} = -x_0^v x_0^t + \langle \mathbf{z}_v, \mathbf{z}_t \rangle, \tag{6}$$

where the mapped feature vectors $\tilde{f}_v = (x_0^v, \mathbf{z}_v) \in \mathbb{R}^{D+1}$, $\tilde{f}_t = (x_0^t, \mathbf{z}_t) \in \mathbb{R}^{D+1}$ and the time components $x_0 = \sqrt{\|\mathbf{z}\|^2 + \kappa^{-1}}$. Due to the time component x_0 being calculated from the spatial component $\mathbf{z}$, the feature vectors $\tilde{f}$ satisfy $\langle \tilde{f}, \tilde{f} \rangle_{\mathbb{H}} = -\kappa^{-1}$. Thus $\tilde{f}$ represents a valid point in $\mathbb{H}^n_\kappa = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{H}} = -\kappa^{-1},\ x_0 > 0\}$ and inherits the geometric properties of hyperbolic space.

(3) **HEMix similarity**: We define the final similarity as a weighted combination of the two:

$$\mathrm{HEMix}(f_v, f_t) = (1-\alpha)\,\mathrm{Sim}_{\mathrm{E}}(f_v, f_t) + \alpha\,\mathrm{Sim}_{\mathrm{H}}(f_v, f_t), \quad \alpha \in (0, 1). \tag{7}$$

Euclidean space ($\kappa = 0$ case) excels at *local, flat* geometry; it preserves fine-grained angular relationships that are crucial for pixel-accurate localization. Lorentzian Hyperbolic space ($\kappa < 0$ case) naturally embeds *hierarchical* or long-range semantics because geodesic distance grows exponentially (Nickel & Kiela, 2018; Ganea et al., 2018), but its metric stretches neighborhoods close to the light cone, which can blur local details. Each geometry similarity ($\mathrm{Sim}_{\mathrm{E}}$ and $\mathrm{Sim}_{\mathrm{H}}$) therefore introduces a different *estimation error* with respect to the ideal but unknown similarity $\mathrm{Sim}^\star$:

$$\mathrm{Sim}_{\mathrm{E}} = \mathrm{Sim}^\star + b_{\mathrm{E}} + \varepsilon_{\mathrm{E}}, \quad \mathrm{Sim}_{\mathrm{H}} = \mathrm{Sim}^\star + b_{\mathrm{H}} + \varepsilon_{\mathrm{H}},$$

Table 1: Comparison on the WGREC task. '†' denotes these methods have been modified to generate multiple boxes following He et al. (2023). '*' denotes the model adapted for the WGREC task.

Methods	Supervision	gRefCOCO						Ref-ZOM	
		val		testA		testB		val	
		Pr (%)	N-acc.(%)	Pr(%)	N-acc.(%)	Pr(%)	N-acc.(%)	Pr(%)	N-acc.(%)
MCN† (Luo et al., 2020)	Fully	28.0	30.6	32.3	32.0	26.8	30.3	-	-
VLT† (Ding et al., 2021)	Fully	36.6	35.2	40.2	34.1	30.2	32.5	-	-
MDETR† (Kamath et al., 2021)	Fully	42.7	36.3	50.0	34.5	36.5	31.0	-	-
UNITEXT† (Yan et al., 2023)	Fully	58.2	50.6	46.4	49.3	42.9	48.2	-	-
Ferret-7B* (You et al., 2023)	Fully	54.8	48.9	49.5	45.2	43.5	43.8	-	-
VistaLLM-7B (Pramanick et al., 2024)	Fully	52.7	69.4	-	-	-	-	-	-
VistaLLM-13B (Pramanick et al., 2024)	Fully	54.6	70.8	-	-	-	-	-	-
RECANTFormer(5) (Hemanthage et al., 2024)	Fully	57.73	52.70	57.82	53.38	49.49	54.53	56.69	88.24
RECANTFormer(10) (Hemanthage et al., 2024)	Fully	55.10	52.73	55.07	53.07	48.01	54.81	59.78	88.24
HieA2G$_{R101}$ (Wang et al., 2025)	Fully	67.8	60.3	66.0	60.1	56.5	56.0	-	-
RefCLIP* (Jin et al., 2023)	Weakly	17.85	0.0	18.23	0.0	21.89	0.0	35.78	0.0
LIHE	**Weakly**	**39.61**	**67.49**	**32.70**	**79.60**	**35.84**	**67.07**	**50.36**	**97.70**

where $b = \mathbb{E}[\text{Sim} - \text{Sim}^\star]$ denotes the *bias* and ε the zero-mean random deviation. Specifically, b_{E} is large for hierarchical descriptions, while b_{H} is large for micro-spatial references. The two errors are not perfectly correlated, due to the characteristics of each space.

Proposition 1 (Variance reduction). *Let* $\sigma_{\text{E}}^2 = \text{Var}[\varepsilon_{\text{E}}]$, $\sigma_{\text{H}}^2 = \text{Var}[\varepsilon_{\text{H}}]$ *and* $\rho = \text{Corr}[\varepsilon_{\text{E}}, \varepsilon_{\text{H}}]$. *If* $\rho < 1$, *the mean-squared error of the hybrid estimator*

$$\begin{aligned}\text{MSE}(\textit{HEMix}) &= \mathbb{E}\big[(\textit{HEMix} - \textit{Sim}^\star)^2\big] \\ &= ((1-\alpha)b_{\text{E}} + \alpha b_{\text{H}})^2 \quad + (1-\alpha)^2\sigma_{\text{E}}^2 + \alpha^2\sigma_{\text{H}}^2 + 2\alpha(1-\alpha)\rho\sigma_{\text{E}}\sigma_{\text{H}},\end{aligned}$$

attains its minimum at $\alpha^\star = \dfrac{(\sigma_{\text{E}}^2 + \rho\sigma_{\text{E}}\sigma_{\text{H}}) + b_{\text{E}}(b_{\text{E}} - b_{\text{H}})}{\sigma_{\text{E}}^2 + \sigma_{\text{H}}^2 - 2\rho\sigma_{\text{E}}\sigma_{\text{H}} + (b_{\text{E}} - b_{\text{H}})^2}$, *and make* MSE(*HEMix*) *satisfies*

$$\text{MSE}(\textit{HEMix}; \alpha^\star) < \min\big\{\text{MSE}(\textit{Sim}_{\text{E}}),\ \text{MSE}(\textit{Sim}_{\text{H}})\big\}. \tag{8}$$

Please see the supplementary material for complete proof. Under the common assumption (Huang et al., 2021b; Gouk et al., 2021; Lei et al., 2023) that the contrastive loss $\mathcal{L}\big(\sigma(S)\big)$ is Lipschitz-continuous in its similarity argument S, a lower MSE implies a tighter generalization bound, which translates into higher retrieval or localization accuracy. Due to ρ is unknown in practice, we select α via experiment. Overall, HEMix unifies two complementary geometric inductive biases, local Euclidean precision and hyperbolic hierarchy, into a single estimator that better approximates $\text{Sim}^\star$, as confirmed empirically in Sec. 4.

4 Experiments

4.1 Datasets and Metrics

We evaluate our proposed method on two benchmark datasets for the WGREC task: gRefCOCO (He et al., 2023) and Ref-ZOM (Wang et al., 2025), both of which support expressions that may correspond to multiple or zero referents. Following prior works (He et al., 2023), we adopt Precision@(F_1=1, IoU$\geq$0.5) and N-acc. as the main evaluation metrics. Precision@(F_1=1, IoU$\geq$0.5) computes the percentage of samples that have the F1 score of 1 with the IoU threshold set to 0.5 and N-acc. assesses the model's proficiency in no-target identification. Detailed explanations are provided in the Appendix F. To further validate the effectiveness and generalization ability of our proposed HEMix module, we also evaluate it on three widely-used WREC datasets: RefCOCO, RefCOCO+, and RefCOCOg. For these datasets, we follow the standard evaluation protocol using IoU@0.5, where a prediction is considered correct if the IoU between the predicted and ground-truth bounding box exceeds 0.5.

4.2 Comparisons with State-of-the-art Methods

WGREC results. As shown in Tab. 1, LIHE achieves strong performance on the gRefCOCO dataset under the weakly supervised setting. Compared to other WREC baselines, RefCLIP (Jin et al., 2023), our model significantly outperforms it across all splits. For instance, on the validation

Table 2: Performance of our methods on RefCOCO, RefCOCO+ and RefCOCOg datasets. '*' indicates results reproduced under identical settings.

Method	Published on	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
VC (Niu et al., 2019)	*CVPR '18*	-	32.68	27.22	-	34.68	28.10	29.65
ARN (Liu et al., 2019c)	*ICCV '19*	32.17	35.25	30.28	32.78	34.35	32.13	33.09
IGN (Zhang et al., 2020)	*NeurIPS '20*	34.78	-	-	36.91	36.91	35.46	34.92
DTWREG (Sun et al., 2021b)	*TPAMI '21*	38.35	39.51	37.01	38.19	39.91	37.09	42.54
RefCLIP* (Jin et al., 2023)	*CVPR '23*	59.88	58.44	56.91	40.11	40.01	38.63	47.87
RefCLIP*+HEMix	-	**60.95**	**59.84**	**58.57**	**41.48**	**42.54**	**39.37**	**48.67**
APL* (Luo et al., 2024)	*ECCV '24*	64.18	61.06	63.08	41.03	41.46	38.72	49.45
APL*+HEMix	-	**65.71**	**62.67**	**64.04**	**42.13**	**42.98**	**40.70**	**50.88**
WeakMCN (Cheng et al., 2025)	*CVPR '25*	69.20	69.88	62.63	51.90	57.33	43.10	54.62
WeakMCN*+HEMix	-	**70.44**	**71.59**	**63.22**	**52.61**	**58.14**	**44.43**	**55.60**

Table 3: Performance of different similarity methods in contrastive loss.

Similarity Method	RefCOCO (WREC)		gRefCOCO (WGREC)		
	testA	testB	val	testA	testB
$\mathrm{Sim_E}$	58.44	56.92	38.88	31.77	34.89
$\mathrm{Sim_H}$	58.94	57.72	39.58	31.57	36.88
HEMix	**59.84**	**58.57**	**39.73**	**32.19**	**37.22**

Table 4: Ablation study on the prompt design with N-acc. metrics

Prompt design	gRefCOCO (WGREC)		
	val	testA	testB
Without P_E	49.00	65.60	50.93
P^*	68.10	76.66	68.26
P	67.49	79.60	67.07

set, our method achieves 39.61% grounding precision and 67.49% normalized accuracy, while RefCLIP only obtains 17.85% and lacks the capability to handle no-target cases. This performance gap clearly demonstrates that methods assuming a single target fail to generalize to WGREC, which involves multi-target and no-target scenarios. Although fully supervised methods like HieA2G (Wang et al., 2025) and RECANTFormer (Hemanthage et al., 2024) outperform our model in absolute metrics, our method remains competitive despite using only image-level supervision, and even surpasses earlier fully supervised baselines such as MCN (Luo et al., 2020), VLT (Ding et al., 2021), and MDETR (Kamath et al., 2021). Additionally, it is worth mentioning that our method can even accomplish unsupervised GREC tasks, as shown in Appendix H.1.

WREC results. We further evaluate the generalization ability of our hybrid similarity learning scheme on standard WREC benchmarks, including RefCOCO, RefCOCO+, and RefCOCOg. As shown in Tab. 2, integrating our proposed hybrid similarity module consistently improves existing baselines. For example, RefCLIP+HEMix surpasses vanilla RefCLIP on all splits (e.g., +1.71% on RefCOCO testA, +1.15% on RefCOCO+testA). Similarly, APL+HEMix yields consistent gains over APL (Luo et al., 2024), improving RefCOCOg by 1.43% and RefCOCO+ testB by 1.98% and WeakMCN+HEMix also improves the performance on three datasets. These results highlight the compatibility and plug-and-play nature of our hybrid similarity formulation. By combining Euclidean and Hyperbolic similarity measures, the model benefits from both local discriminative alignment and global semantic consistency, leading to better grounding performance across varying datasets and expression types.

4.3 Ablation Study

To validate the effectiveness of individual components in our framework, we conduct comprehensive ablation studies covering similarity modules, mapping strategies and cross-dataset validation.

Effect of Similarity Design in Contrastive Learning. We conduct an ablation study to assess the impact of different similarity functions in contrastive learning, as shown in Tab. 3. Overall, our proposed HEMix consistently outperforms standard Euclidean similarity ($\mathrm{Sim_E}$, equivalent to Referent Grounding directly using RefCLIP), with average gains of 1.53% on RefCOCO (WREC) and 0.90% on gRefCOCO (WGREC), demonstrating its effectiveness across both tasks. When comparing SimH and SimE, the improvement is more pronounced on WGREC than on WREC (+0.83% vs. +0.65%), and notably, $\mathrm{Sim_H}$ achieves performance close to HEMix on WGREC, with only a 0.44% gap on average. This suggests that hyperbolic similarity better supports the grounding of multiple semantically related referents, which is common in WGREC. One likely reason is that expressions in WGREC often correspond to multiple related but distinct instances, requiring a more structured

Table 5: Cross-Dataset validation from GREC to REC benchmarks. In-Dataset denotes the model trained on the same dataset as the test.

Method	Training Set	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
RefCLIP	In-Dataset	59.88	58.44	56.91	40.11	40.01	38.63	47.87
Ours	In-Dataset	**60.95**	**59.84**	**58.57**	41.48	42.54	**39.37**	48.67
RefCLIP	gRefCOCO	47.62	46.83	48.53	38.25	38.91	38.00	43.73
Ours	gRefCOCO	54.34	55.79	51.38	**42.04**	**43.22**	38.84	**49.14**

Table 6: Performance of different hyperbolic mapping methods on RefCOCO and WGREC.

Hyperbolic Mapping Method	RefCOCO			WGREC		
	val	testA	testB	val	testA	testB
Exponential Map	55.02	54.16	53.78	38.65	31.44	34.37
Learnable Linear	**60.95**	**59.84**	**58.57**	**39.61**	**32.70**	**35.84**

representation space, an advantage naturally offered by hyperbolic geometry. These findings further confirm the suitability of hyperbolic space for modeling hierarchical semantics. Nevertheless, HEMix consistently outperforms both SimE and SimH across all datasets. Even in cases where SimE and SimH perform similarly (e.g., RefCOCO testA, gRefCOCO testA), HEMix achieves great improvements. This highlights the importance of incorporating both fine-grained and hierarchical information and demonstrates the overall superiority of our proposed HEMix.

Ablation Study on Prompt design We conduct an ablation study on prompt design and use N-acc as metrics. The prompt P contains P_G, P_C, P_E, P_Q and I. Among them, P_G, P_C, P_Q, and I are indispensable; the absence of any one would lead to unparseable VLM output and consequently task failure. Therefore, we present two variants: one with P_E removed, and another, P^*, where the textual content is altered while preserving the semantic information of P.Removing P_E led to some outputs not following the format well and reduced the model's understanding of the task, which in turn led to performance degradation. The performance of P^* when changing the text content was similar to P, indicating the robustness of our method to prompts.

Effect of Different Hyperbolic Mapping Designs. Prior works (Gao et al., 2021; Khrulkov et al., 2020; Desai et al., 2023) typically adopt the exponential map to project features onto the hyperboloid manifold, preserving the mathematical correctness of hyperbolic embeddings. However, we observe that this mapping introduces steep gradients, which negatively affect training stability and optimization. As shown in Tab. 6, replacing the exponential map with a learnable linear layer yields a substantial improvement of 5.24% on RefCOCO and 1.23% on gRefCOCO, indicating that a simpler, trainable mapping leads to better empirical performance in practice. To our best knowledge, only these two hyperbolic mapping schemes exist.

Cross-Dataset Validation. To evaluate the robustness and generalization of our framework, we train the model on the gRefCOCO and test it on standard WREC benchmarks with single-target annotations. As shown in Tab. 5, our method consistently outperforms the single-target baseline RefCLIP across all test sets, despite being trained on the same gRefCOCO data. Compared to in-domain training, the cross-dataset setting yields even larger gains (e.g., +5.41% *vs.* +0.80% in RefCOCOg), further highlighting the superior transferability of our approach. Notably, our model even surpasses in-domain RefCLIP models on RefCOCO+ and RefCOCOg, demonstrating strong generalization across both task settings and dataset domains.

Furthermore, our more ablation study in Appendix H.2 reveals several consistent trends across datasets. First, sweeping the hybrid weight α for HEMix produces a clear U-shaped curve with a robust sweet spot around $\alpha \in [0.4, 0.7]$; for example, RefCOCO testA peaks at 60.01% when α=0.7, while gRefCOCO testB reaches 36.44% at α=0.9 (Tab. 11). Second, adding an *explicit* hierarchical constraint brings little to no average improvement over the *implicit* structure already captured by HEMix (Tab. 13). Third, during referent grounding, excluding v=0 samples avoids degenerate contrastive updates and improves performance, while such cases are handled at inference by the referential decoupling stage, outputting "0" (Tab. 12).

4.4 Quantitative Analysis

To further understand the behavior of our model in complex referring scenarios, we visualize representative success and failure cases from the gRefCOCO dataset in Fig. 4. These examples highlight the strengths and current limitations of our framework.

Figure 4: Successful cases(green background) and failure cases(red background). The ground truth is denoted by red bounding boxes, whereas green bounding boxes denote the predictions.

Success Cases. As shown in Fig. 4(A), our method demonstrates strong grounding ability across both multi-target and no-target expressions. For instance, in case (a), the expression includes multiple entities with spatial and appearance constraints, and our model is able to localize each person correctly despite heavy occlusion and crowd density. In (e), although the objects (donuts) are visually similar, our model grounds the correct ones by leveraging position cues. Moreover, our approach can successfully identify no-target cases (c) and (f). In (c), the phrase *"the leftmost person wearing black uniform"* does not correspond to any entity in the image, and our model makes a correct no-target prediction. In (f), while the scene contains many people, none of them match the detailed attributes described in expression (*"woman in red dress"*), and the model again avoids false positives.

Failure Cases. Fig. 4(B) presents typical failure cases. In (a), the expression implies no valid target, yet the model incorrectly grounds an entity, reflecting the limited semantic understanding of the VLM. In (b), the number of targets is predicted correctly, but all decomposed expressions collapse to the same label ("people"), caused by VLM hallucinations. This occurs despite correctly detecting four objects (two balls and two rackets); class imbalance biases the decoder toward the frequent category "people," consistent with prior findings on distribution-induced hallucination (Zhang et al., 2024; McKenna et al., 2023; Rohrbach et al., 2018; Liang et al., 2025). In (c), although the decoupled phrases are semantically valid and distinct, the grounding stage fails to segment precise regions, leading to overlaps and missed detections. In (d), the model misses fine-scale visual cues, such as small hands or blurred individuals, demonstrating difficulty in handling subtle details.

5 Conclusion

In this paper, we delve into the critical limitation of Weakly-Supervised Referring Expression Comprehension (WREC) task: the inability of existing methods to handle expressions corresponding to a variable number of targets. To address this, we introduce the Weakly-Supervised Generalized Referring Expression Comprehension (WGREC) task, a more realistic and challenging setting where an expression may refer to multiple, single, or no objects. We then propose LIHE, the first weakly-supervised framework designed for this generalized task. LIHE operates via a two-stage process: (1) Referential Decoupling, where a vision-language model (VLM) infers the number of potential referents and parses the expression into target-relevant sub-phrases, followed by (2) Referent Grounding, where the model enhanced by our novel hybrid similarity localizes each sub-phrase. This design leverages the semantic understanding of VLMs to resolve ambiguity and synergistically combines Euclidean and hyperbolic geometries to preserve hierarchical representations. Despite these strengths, experiments demonstrate that LIHE achieves state-of-the-art performance on both WGREC and WREC benchmarks, validating its robustness and generalization. One key limitation of LIHE is its reliance on VLMs, which limits inference speed ($\approx$2 FPS). Thus, LIHE is more suitable as a teacher model for generating pseudo-labels (Jiang et al., 2024) to supervise smaller, faster student models. In future work, we will explore student model design and lightweight VLM adaptation, as well as extend our hybrid similarity design to richer geometric formulations.

Reproducibility Statement

All theoretical results are established under explicit conditions. Experimental details are given in Appendix C. The code and weights are available at https://anonymous.4open.science/r/LIHE.

Ethics Statement

This work only uses publicly available datasets and does not involve human subjects or sensitive information. We identify no specific ethical concerns.

References

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4453–4462, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021.

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

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16321–16330, 2021.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7409–7419, 2022.

Octavian-Emanuel Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, pp. 5345–5355, 2018.

Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Curvature generation in curved spaces for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8691–8700, 2021.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6840–6849, 2023.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pp. 752–768, 2020.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023.

Bhathiya Hemanthage, Hakan Bilen, Phil Bartie, Christian Dondrup, and Oliver Lemon. Recantformer: Referring expression comprehension with varying numbers of targets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21784–21798, 2024.

Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. Yoro—lightweight end to end visual grounding. In *European Conference on Computer Vision*, pp. 3–23, 2022.

Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):684–696, 2019.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4067–4077, 2023.

Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16888–16897, 2021a.

Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021b.

Yunhan Jiang, Xianglong Shi, Xiaoheng Jiang, Jian Feng, Yang Lu, and Mingliang Xu. Diffusaliency: Synthesizing multi-object images with masks for semantic segmentation using diffusion and saliency detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 74–88. Springer, 2024.

Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2681–2690, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6418–6428, 2020.

Sungyeon Kim, Boseung Jeong, and Suha Kwak. Hier: Metric learning beyond class labels via hierarchical regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19903–19912, 2023.

Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. Hyperbolic learning with synthetic captions for open-world detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16762–16771, 2024.

Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. Improving visual recognition with hyperbolical visual hierarchy mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17364–17374, 2024.

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *International Conference on Machine Learning*, pp. 19200–19227. PMLR, 2023.

Huimin Li, Zhentao Chen, Yunhao Xu, and Junlin Hu. Hyperbolic anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17511–17520, 2024.

Xiaoyu Liang, Jiayuan Yu, Lianrui Mu, Jiedong Zhuang, Jiaqi Hu, Yuchen Yang, Jiangnan Ye, Lu Lu, Jian Chen, and Haoji Hu. Mitigating hallucination in visual-language models via rebalancing contrastive decoding. In *Pattern Recognition and Computer Vision (PRCV 2024)*, volume 15035 of *Lecture Notes in Computer Science*, pp. 482–496, Singapore, 2025. Springer. doi: 10.1007/978-981-97-8620-6_33. URL https://link.springer.com/chapter/10.1007/978-981-97-8620-6_33.

Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10880–10889, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23592–23601, 2023.

Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4673–4682, 2019a.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1950–1959, 2019b.

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

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

Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5612–5621, 2021.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 10034–10043, 2020.

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2758–2774, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.182. URL https://aclanthology.org/2023.findings-emnlp.182/.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 792–807. Springer, 2016.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, pp. 6338–6347, 2017a.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017b.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pp. 3779–3788. PMLR, 2018.

Yulei Niu, Hanwang Zhang, Zhiwu Lu, and Shih-Fu Chang. Variational context: Exploiting visual and textual context for grounding referring expressions. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):347–359, 2019.

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

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14076–14088, 2024.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27263–27272, 2024.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Xianglong Shi, Yunhan Jiang, Xiaoheng Jiang, Mingling Xu, and Yang Liu. Crossdiff: Diffusion probabilistic model with cross-conditional encoder-decoder for crack segmentation. *arXiv preprint arXiv:2501.12860*, 2025.

Mingjie Sun, Jimin Xiao, and Eng Gee Lim. Iterative shrinking for referring expression grounding using deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14060–14069, 2021a.

Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4189–4195, 2021b.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14090–14100, 2021.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yaxian Wang, Henghui Ding, Shuting He, Xudong Jiang, Bifan Wei, and Jun Liu. Hierarchical alignment-enhanced adaptive grounding network for generalized referring expression comprehension. *arXiv preprint arXiv:2501.01416*, 2025.

Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, and Serena Yeung. Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2603–2612, 2021.

Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun, and Rongrong Ji. 3d-gres: Generalized 3d referring expression segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7852–7861, 2024.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15325–15336, 2023.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4683–4693, 2019.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision—ECCV 2020*, pp. 387–404, 2020.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4158–4166, 2018.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R. Fung, Jing Li, Manling Li, and Heng Ji. Knowledge overshadowing causes amalgamated hallucination in large language models, 2024. URL https://arxiv.org/abs/2407.08039.

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

Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5696–5705, 2018.

Heng Zhao, Joey Tianyi Zhou, and Yew-Soon Ong. Word2pix: Word to pixel cross-attention transformer in visual grounding. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (2):1523–1533, 2022.

Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):134–143, 2021a.

Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2074–2084, 2021b.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pp. 598–615. Springer, 2022.

Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. Hypertext: Endowing fasttext with hyperbolic geometry. *arXiv preprint arXiv:2010.16143*, 2020.

APPENDIX

A USE OF LARGE LANGUAGE MODELS

We use large language models to aid or polish writing.

B RELATED WORK

Generalized Referring Expression Comprehension GREC (He et al., 2023) extends traditional REC tasks by permitting linguistic expressions to refer simultaneously to multiple target objects. Baseline methods such as MCN (Luo et al., 2020), VLT (Ding et al., 2021), MDETR (Kamath et al., 2021), and UNINEXT (Yan et al., 2023) have been adopted to evaluate performance under this more complex and realistic scenario. Recent advancements include RECANTFormer (Hemanthage et al., 2024), which employs a recursive transformer decoder with adaptive prediction heads to dynamically predict multiple targets, and HieA2G (Wang et al., 2025), which leverages a hierarchical alignment mechanism to enhance interactions between linguistic phrases and visual objects, thereby capturing fine-grained semantic correlations. However, these methods uniformly rely on fully supervised bounding box annotations. Weakly supervised approaches for GREC remain unexplored, motivating the development of our proposed weakly supervised framework.

Weakly Supervised REC Weakly supervised approaches (Gupta et al., 2020; Liu et al., 2019c;d; 2021; Sun et al., 2021b; Wang et al., 2021; Zhang et al., 2020) have recently shown promising results in Referring Expression Comprehension (REC), significantly reducing the dependence on expensive bounding box annotations. Unlike fully supervised methods (Deng et al., 2021; Ho et al., 2022; Hong et al., 2019; Huang et al., 2021a; Kamath et al., 2021; Liao et al., 2020; Liu et al., 2019a;b; Shi et al., 2025; Luo et al., 2020; Sun et al., 2021a; Yang et al., 2019; 2020; Zhang et al., 2018; Zhao et al., 2022; Zhou et al., 2021a;b; Zhu et al., 2022), weakly supervised REC techniques utilize coarser supervision signals, such as image-level or image-text pair annotations, enabling scalable and cost-effective training. Notable one-stage methods (Jin et al., 2023; Luo et al., 2024; Zhao et al., 2018) include RefCLIP (Jin et al., 2023), which redefines REC as an anchor-text matching task using anchor-based contrastive learning to align visual and textual features without bounding boxes, and APL (Luo et al., 2024), which enriches anchor features through position, color, and category prompts, coupled with auxiliary losses to enhance vision-language alignment. Despite their effectiveness in traditional REC tasks, these methods inherently assume a single-target scenario, limiting their generalizability to multi-target WGREC settings. This highlights the need to develop specialized, weakly supervised approaches tailored specifically to GREC.

Hyperbolic Representation learning Hyperbolic representation learning (Tifrea et al., 2018; Zhu et al., 2020; Nickel & Kiela, 2018; 2017b; Kim et al., 2023; Ermolov et al., 2022; Atigh et al., 2022; Weng et al., 2021; Gao et al., 2021; Khrulkov et al., 2020; Desai et al., 2023) leverages hyperbolic geometry's exponential volume growth and negative curvature, making it particularly effective for modeling hierarchical and relational structures. Early seminal works such as Poincaré Embeddings (Nickel & Kiela, 2017a) and Hyperbolic Neural Networks (Ganea et al., 2018) established foundational techniques for embedding structured data into hyperbolic spaces. Recent developments have adapted hyperbolic geometry to vision and cross-modal tasks. For instance, hyperbolic embedding methods introduced by Kwon et al. (Kwon et al., 2024) and Kong et al. (Kong et al., 2024) have improved visual recognition and open-world detection by preserving hierarchical semantics. For vision-language alignment, Ge et al. (Ge et al., 2023) and Ramasinghe et al. (Ramasinghe et al., 2024) used hyperbolic contrastive learning to enhance semantic coherence. Unlike previous methods that use only hyperbolic distance or angle, we fuse Euclidean and hyperbolic similarities into a hybrid metric, which boosts WREC performance and showcases a new way to apply hyperbolic geometry in representation learning.

C IMPLEMENTATION DETAILS

In the referential decoupling stage, we adopt a pre-trained VLM (Team, 2025) to understand the visual entity and generate the decomposed referring expression. In the referent grounding stage, following prior work, we resize every image to 416×416. The maximum token length of the referring expression is fixed to 15 for all datasets. For anchor extraction, we adopt the YOLOv3 (Redmon &

Farhadi, 2018) model pre-trained on MSCOCO (Lin et al., 2014), where images from the validation and test splits of those datasets are removed to avoid leakage. Detector weights are frozen during all stages of training. The language encoder produces 512-dimensional sentence embeddings. Visual anchor features are first fused across scales and then linearly projected to the same 512-dimensional joint space. In anchor-based contrastive learning, we adopt a 512-dimensional projection head and sample two negative anchors per image by default. All WREC tasks are trained on NVIDIA GPU A100 40G and all WGREC tasks on A6000 48G. All models are optimized with AdamW using a constant learning rate of $1e-4$. We train for 25 epochs with a batch size of 64.

D Hyperbolic Space Properties

Hyperbolic spaces are Riemannian manifolds characterized by negative curvature, and they differ fundamentally depending on the curvature value. As the curvature approaches zero, the hyperbolic space gradually transitions into a Euclidean space. When the curvature is negative, the space exhibits hyperbolic geometry, where parallel lines can diverge, and the volume grows exponentially with distance from a point. These spaces are commonly represented using various models, including the Poincaré ball, the Lorentz model, and the Klein model, each providing unique advantages for mathematical formulations.

Lorentz Model The Lorentz Model is also known as the hyperboloid model or the Minkowski model. In the Lorentz model, a hyperbolic n-dimensional manifold is commonly realized as a sub-manifold of $\mathbb{R}^{n+1}$, corresponding to the upper sheet of a two-sheeted hyperboloid. Each point $\mathbf{x} \in \mathbb{R}^{n+1}$ of the Lorentz model, can be represented as $[\mathbf{x}_{\text{time}}, x_{\text{space}}]$, where $x_{\text{time}} \in \mathbb{R}$ denotes the temporal component and $\mathbf{x}_{\text{space}} \in \mathbb{R}^n$ denotes the spatial components. The n-dimensional Hyperbolic space $\mathbb{H}^n$ with curvature κ represented by a $n+1$-dimensional Lorentz Model as follows:

$$\mathbb{H}^n_\kappa \;=\; \Big\{\mathbf{x} \in \mathbb{R}^{n+1} \;\Big|\; \langle \mathbf{x}, \mathbf{x}\rangle_{\mathbb{H}} \;=\; -\kappa^{-1},\; x_0 > 0\Big\} \tag{9}$$

where the Lorentzian inner product is defined as,

$$\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{H}} \;=\; -x_0 y_0 \;+\; \sum_{i=1}^{n} x_i y_i. \tag{10}$$

Here, the 0-th dimension of the vector $\mathbf{x}$, x_0 is treated as the time component x_{time} and the rest dimension of the vector $\mathbf{x}$, $\mathbf{x}_{1:n}$ is the space component $\mathbf{x}_{space}$. The x_{time} can be calculated from $\mathbf{x}_{space}$ as follows:

$$x_{time} = x_0 = \sqrt{\|\mathbf{x}_{\text{space}}\|^2 + \kappa^{-1}}\,. \tag{11}$$

where the $\|\ldots\|$ is the Euclidean norm.

Distance In the Lorentz model, the geodesic distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^n_\kappa$ ($\kappa > 0$) can be expressed solely in terms of their Lorentzian inner product. The distance function $d_\kappa : \mathbb{H}^n_\kappa \times \mathbb{H}^n_\kappa \to \mathbb{R}_{\geq 0}$ is

$$d_\kappa(\mathbf{x}, \mathbf{y}) \;=\; \frac{1}{\sqrt{\kappa}}\,\operatorname{arccosh}\Big(-\kappa\,\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{H}}\Big), \tag{12}$$

where $\operatorname{arccosh}$ is the inverse hyperbolic cosine and $\langle\cdot,\cdot\rangle_{\mathbb{H}}$ is lorentz inner product in equation 10. The quantity inside the $\operatorname{arccosh}$ is always ≥ 1 for points on the hyperboloid, guaranteeing that the distance is real-valued. This formula highlights a key feature of hyperbolic geometry: the distance grows logarithmically with the Lorentzian inner product, reflecting the exponential volume growth characteristic of negatively curved spaces.

Exponential Map Given a base point $\mathbf{p} \in \mathbb{H}^n_\kappa$ and a tangent vector $\mathbf{v} \in T_{\mathbf{p}}\mathbb{H}^n_\kappa$ (Euclidean space), the *exponential map* $\operatorname{Exp}^\kappa_{\mathbf{p}} : T_{\mathbf{p}}\mathbb{H}^n_\kappa \to \mathbb{H}^n_\kappa$ moves $\mathbf{p}$ along the unique geodesic in the direction of $\mathbf{v}$. the map is

$$\operatorname{Exp}^\kappa_{\mathbf{p}}(\mathbf{v}) \;=\; \cosh\!\big(\sqrt{\kappa}\,\|\mathbf{v}\|\big)\,\mathbf{p} \;+\; \frac{\sinh\!\big(\sqrt{\kappa}\,\|\mathbf{v}\|\big)}{\sqrt{\kappa}\,\|\mathbf{v}\|}\,\mathbf{v}, \tag{13}$$

where $\|\cdots\|$ for the Lorentz norm o, For small $\|\mathbf{v}\|$ this reduces to $\mathbf{p}+\mathbf{v}$, mirroring the Euclidean limit, while for large $\|\mathbf{v}\|$ the hyperbolic $\cosh$ / $\sinh$ terms dominate, capturing the curvature-induced stretching of space.

Logarithm Map Conversely, the *logarithm map* $\mathrm{Log}_{\mathbf{p}}^{\kappa} : \mathbb{H}_{\kappa}^{n} \to T_{\mathbf{p}}\mathbb{H}_{\kappa}^{n}$ sends a point $\mathbf{q}$ back to the tangent space at $\mathbf{p}$, producing the initial velocity vector of the geodesic from $\mathbf{p}$ to $\mathbf{q}$. Using the distance $d_{\kappa}(\mathbf{p}, \mathbf{q})$ from equation 12, we have

$$\mathrm{Log}_{\mathbf{p}}^{\kappa}(\mathbf{q}) \;=\; \frac{d_{\kappa}(\mathbf{p}, \mathbf{q})}{\sinh\big(\sqrt{\kappa}\, d_{\kappa}(\mathbf{p}, \mathbf{q})\big)} \left(\mathbf{q} + \kappa \, \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{H}} \, \mathbf{p}\right), \tag{14}$$

which indeed satisfies $\mathrm{Exp}_{\mathbf{p}}^{\kappa}\big(\mathrm{Log}_{\mathbf{p}}^{\kappa}(\mathbf{q})\big) = \mathbf{q}$. The factor in front rescales the component of $\mathbf{q}$ orthogonal to $\mathbf{p}$ so that its norm equals the hyperbolic distance, giving a first-order approximation to motion on the manifold that is exact along geodesics.

From these definitions, we highlight two key insights:**(1) Any vector that satisfies the hyperboloid constraint (Eq. equation 9) is a valid point on the Lorentz manifold and inherits the geometric properties of hyperbolic space.(2) A higher Lorentzian inner product indicates greater semantic similarity between two points in hyperbolic space.**

Proposition 2 (Hyperboloid Membership)**.** *Let $\kappa > 0$. A vector $\mathbf{x} \in \mathbb{R}^{n+1}$ with $x_0 > 0$ satisfies the hyperboloid constraint*

$$\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{H}} = -\kappa^{-1} \quad \Longleftrightarrow \quad \mathbf{x} \in \mathbb{H}_{\kappa}^{n}.$$

Consequently every such $\mathbf{x}$ is a valid point of the Lorentz (hyperboloid) model of the hyperbolic space of constant sectional curvature $-\kappa$, and inherits all of its geometric properties.

Proof. *($\Rightarrow$)* By definition

$$\mathbb{H}_{\kappa}^{n} \;=\; \Big\{ \mathbf{z} \in \mathbb{R}^{n+1} \;\Big|\; \langle \mathbf{z}, \mathbf{z} \rangle_{\mathbb{H}} = -\kappa^{-1},\; z_0 > 0 \Big\},$$

so any $\mathbf{x}$ satisfying the stated constraint (with $x_0 > 0$) belongs to $\mathbb{H}_{\kappa}^{n}$.

($\Leftarrow$) Conversely, if $\mathbf{x} \in \mathbb{H}_{\kappa}^{n}$, then by the same defining condition we have $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{H}} = -\kappa^{-1}$ and $x_0 > 0$. Hence the two sets coincide, establishing the equivalence. □

Proposition 3 (Monotonicity of the Lorentzian Inner Product)**.** *For any two points $\mathbf{x}, \mathbf{y} \in \mathbb{H}_{\kappa}^{n}$ with $\kappa > 0$, the geodesic distance equation 12*

$$d_{\kappa}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\kappa}} \, \mathrm{arccosh}\big(-\kappa \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}\big)$$

is a strictly decreasing *function of the Lorentzian inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}$. Equivalently, a **larger inner product indicates smaller hyperbolic distance** and therefore higher semantic similarity.*

Proof. Define

$$u \;:=\; -\kappa \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}} \quad \big(\, u \geq 1 \,\big)$$

$$f(u) \;:=\; \frac{1}{\sqrt{\kappa}} \, \mathrm{arccosh}(u) = \; d_{\kappa}(\mathbf{x}, \mathbf{y}).$$

Because $\mathrm{arccosh}$ is strictly increasing on $[1, \infty)$ and

$$\frac{du}{d\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}} = -\kappa < 0,$$

the chain rule gives

$$\frac{d\, d_{\kappa}(\mathbf{x}, \mathbf{y})}{d\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}} = \frac{df}{du} \, \frac{du}{d\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}} < 0.$$

Hence d_{κ} decreases strictly as $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}$ increases. Since hyperbolic distance quantifies dissimilarity, the inverse relationship asserts that a larger (i.e., less negative) Lorentzian inner product encodes greater semantic similarity between the points. □

Hence, we adopt the Lorentz inner product as a similarity measure.

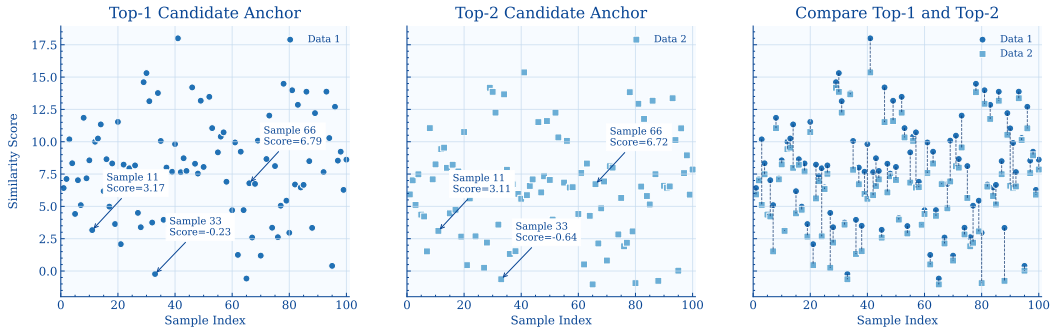


Figure 5: We randomly select 100 samples from the dataset and visualize their similarity scores. From left to right: (1) the similarity scores of top-scoring anchors (dark blue dots), (2) the similarity scores of second-best anchors (light blue squares), and (3) an overlaid view combining both. The dashed vertical segments connect the top and second-best scores for each sample, illustrating that second-best anchors in many cases have higher similarity than top anchors in other samples.

Proposition 4 (Bias–Variance Reduction). *Let* $\sigma_{\mathrm{E}}^2 = \mathrm{Var}[\varepsilon_{\mathrm{E}}]$, $\sigma_{\mathrm{H}}^2 = \mathrm{Var}[\varepsilon_{\mathrm{H}}]$, *and* $\rho = \mathrm{Corr}[\varepsilon_{\mathrm{E}}, \varepsilon_{\mathrm{H}}]$. *Denote the Euclidean and Hyperbolic biases by* $b_{\mathrm{E}} = \mathbb{E}[\mathit{Sim}_{\mathrm{E}} - \mathit{Sim}^\star]$ *and* $b_{\mathrm{H}} = \mathbb{E}[\mathit{Sim}_{\mathrm{H}} - \mathit{Sim}^\star]$. *For any mixing weight* $\alpha \in [0,1]$ *define the hybrid estimator HEMix* $= (1-\alpha)\mathit{Sim}_{\mathrm{E}} + \alpha \mathit{Sim}_{\mathrm{H}}$.

Then the mean–squared error (MSE) of HEMix

$$\begin{aligned}\mathrm{MSE}(\mathit{HEMix}) &= \mathbb{E}\big[(\mathit{HEMix} - \mathit{Sim}^\star)^2\big] \\ &= \big((1-\alpha)b_{\mathrm{E}} + \alpha b_{\mathrm{H}}\big)^2 + (1-\alpha)^2\sigma_{\mathrm{E}}^2 + \alpha^2\sigma_{\mathrm{H}}^2 + 2\alpha(1-\alpha)\rho\sigma_{\mathrm{E}}\sigma_{\mathrm{H}} \qquad (15)\end{aligned}$$

is a strictly convex quadratic in α. *Its unique minimizer is*

$$\alpha^\star = \frac{(\sigma_{\mathrm{E}}^2 + \rho\sigma_{\mathrm{E}}\sigma_{\mathrm{H}}) + b_{\mathrm{E}}(b_{\mathrm{E}} - b_{\mathrm{H}})}{\sigma_{\mathrm{E}}^2 + \sigma_{\mathrm{H}}^2 - 2\rho\sigma_{\mathrm{E}}\sigma_{\mathrm{H}} + (b_{\mathrm{E}} - b_{\mathrm{H}})^2}, \qquad (16)$$

which always lies in $(0,1)$ *whenever* $\rho < 1$ *or* $b_{\mathrm{E}} \neq b_{\mathrm{H}}$. *Moreover,*

$$\mathrm{MSE}\big(\mathit{HEMix}; \alpha^\star\big) \;<\; \min\big\{\mathrm{MSE}(\mathit{Sim}_{\mathrm{E}}),\ \mathrm{MSE}(\mathit{Sim}_{\mathrm{H}})\big\}. \qquad (17)$$

Proof. Let $f(\alpha) = \mathrm{MSE}(\mathrm{HEMix})$ in equation 15. Write it as $f(\alpha) = A\alpha^2 + 2B\alpha + C$ with

$$\begin{aligned}A &= (b_{\mathrm{H}} - b_{\mathrm{E}})^2 + \sigma_{\mathrm{H}}^2 + \sigma_{\mathrm{E}}^2 - 2\rho\sigma_{\mathrm{E}}\sigma_{\mathrm{H}} \;>\; 0, \\ B &= -\big[(b_{\mathrm{H}} - b_{\mathrm{E}})b_{\mathrm{E}} + \sigma_{\mathrm{H}}^2 - \rho\sigma_{\mathrm{E}}\sigma_{\mathrm{H}}\big], \\ C &= b_{\mathrm{E}}^2 + \sigma_{\mathrm{E}}^2 = f(0).\end{aligned}$$

Since $A > 0$, f is strictly convex; the stationary point $\alpha^\star = -B/A$ is the global minimum, yielding $f(\alpha^\star) = C - B^2/A$. Because $B^2/A > 0$, we have $f(\alpha^\star) < f(0) = \mathrm{MSE}(\mathrm{Sim}_{\mathrm{E}})$. Convexity further implies $f(\alpha^\star) < \max\{f(0), f(1)\}$. Whenever $f(1) \neq f(0)$ (i.e., the two single-space estimators do not have identical MSE) this gives $f(\alpha^\star) < \min\{f(0), f(1)\}$, which is exactly the desired inequality. □

E Detailed Limitations of RefCLIP

As the main text says, weakly supervised methods such as RefCLIP effectively simplify the REC task by reducing it to an anchor-text matching problem. Specifically, the anchor selection mechanism in methods like RefCLIP can be expressed as:

$$a^* = \arg\max_{a \in \mathcal{A}} \phi(T, I, a), \qquad (18)$$

where $\phi(T, I, a)$ represents the similarity between the text expression T, the image I, and anchor a from the anchor set $\mathcal{A}$. However, this max-selection strategy implicitly assumes that the number

of referred objects is known in advance, making it unsuitable for Generalized Referring Expression Comprehension (GREC), where the number of targets is unknown.

A natural alternative is to apply a threshold to filter anchors based on their similarity scores. Yet, our experimental analysis reveals the limitations of this approach. As shown in Fig. 5, the second-best anchor in some samples exhibits higher similarity than the top-scoring anchors in others, rendering a universal threshold ineffective for consistent selection. This highlights the inadequacy of threshold-based selection under WGREC conditions.

F Evaluation Metrics

Precision@(F_1=1, IoU$\geq$0.5) For each sample, let $\mathcal{G}$ and $\mathcal{P}$ be the ground-truth and predicted bounding-box sets. A prediction is **matched** to a ground-truth box if their intersection-over-union (IoU) is at least 0.5. If several predictions match the *same* ground-truth box, only the one with the highest IoU is kept as a true positive (TP); the rest are false positives (FP). Unmatched ground-truth boxes are false negatives (FN). The sample-level F_1 score is

$$F_1 = \frac{2\,TP}{2\,TP + FP + FN}.$$

For *no-target* samples ($|\mathcal{G}| = 0$), we set $F_1 = 1$ if $|\mathcal{P}| = 0$ and $F_1 = 0$ otherwise.

Precision@(F_1=1, IoU$\geq$0.5) is the proportion of samples whose F_1 score equals 1:

$$\text{Precision@}(F_1{=}1,\ IoU{\geq}0.5) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[F_1^i = 1\right].$$

It reports the percentage of images perfectly predicted (no missed or spurious detections) under the 0.5 IoU criterion.

N-acc. quantifies the model's ability to correctly identify *no-target* samples—images that contain no ground-truth objects. For such a sample,

- **true positive (TP):** the model predicts *no* bounding boxes;
- **false negative (FN):** the model predicts at least one bounding box.

The metric is

$$\text{N-acc.} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

the proportion of no-target samples for which the model outputs no detections.

G Detailed Prompt Design

$\mathbf{P_G}$: Task Explanation: You need to process an image and a referring expression. The image may contain zero, one, or multiple target objects corresponding to the referring expression. Analyze the image to determine whether the target exists. If the target does not exist or the referring expression is empty, output a single number "0". If the target exists, output the number of targets and generate a unique referring expression for each target. The referring expressions must describe distinct targets unambiguously using attributes like color, position, size, etc.

$\mathbf{P_C}$: You should provide a number indicating how many targets exist in the image, and then describe each target with a short, distinct phrase. Prefix each phrase with its ordinal number. The number of targets is extremely important — please check carefully. The phrases must be accurate and distinct.

$\mathbf{P_E}$: For example, if the referring expression is "3 people", you should output: "3\n1. person ...\n2. person ...\n3. person ..." The word "and" is generally used between two target items.

$\mathbf{P_Q}$: The referring expression is: {referring expression}

The specific usage process of prompts are shown in Tab. 7, Tab. 8 and Tab. 9:

System: Task Explanation: You need to process an image and a referring expression. The image may contain zero, one, or multiple target objects corresponding to the referring expression. Analyze the image to determine whether the target exists. If the target does not exist or the referring expression is empty, output a single number ”0”. If the target exists, output the number of targets and generate a unique referring expression for each target. The referring expressions must describe distinct targets unambiguously using attributes like color, position, size, etc.
User: The referring expression is: {the right boy in black shirt is playing skateboard}. You should provide a number indicating how many targets exist in the image, and then describe each target with a short, distinct phrase. Prefix each phrase with its ordinal number. The number of targets is extremely important — please check carefully. The phrases must be accurate and distinct. For example, if the referring expression is “3 people”, you should output: “3\n1. person ...\n2. person ...\n3. person ...”The word “and” is generally used between two target items.
Referential Decoupling: “0”
No Referent Grounding and directly Return No-target

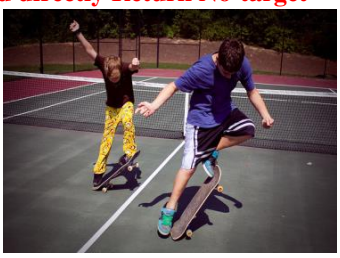


Table 7: No-target case.

System: Task Explanation: You need to process an image and a referring expression. The image may contain zero, one, or multiple target objects corresponding to the referring expression. Analyze the image to determine whether the target exists. If the target does not exist or the referring expression is empty, output a single number ”0”. If the target exists, output the number of targets and generate a unique referring expression for each target. The referring expressions must describe distinct targets unambiguously using attributes like color, position, size, etc.
User: The referring expression is: {a guy in green and a rightmost guy}. You should provide a number indicating how many targets exist in the image, and then describe each target with a short, distinct phrase. Prefix each phrase with its ordinal number. The number of targets is extremely important — please check carefully. The phrases must be accurate and distinct. For example, if the referring expression is “3 people”, you should output: “3\n1. person ...\n2. person ...\n3. person ...”The word “and” is generally used between two target items.
Referential Decoupling: “2\n1.a guy in green\n2.a rightmost guy”
Referent Grounding:

Table 8: One-target case.

System: Task Explanation: You need to process an image and a referring expression. The image may contain zero, one, or multiple target objects corresponding to the referring expression. Analyze the image to determine whether the target exists. If the target does not exist or the referring expression is empty, output a single number ”0”. If the target exists, output the number of targets and generate a unique referring expression for each target. The referring expressions must describe distinct targets unambiguously using attributes like color, position, size, etc.
User: The referring expression is: {three glasses}. You should provide a number indicating how many targets exist in the image, and then describe each target with a short, distinct phrase. Prefix each phrase with its ordinal number. The number of targets is extremely important — please check carefully. The phrases must be accurate and distinct. For example, if the referring expression is “3 people”, you should output: “3\n1. person ...\n2. person ...\n3. person ...”The word “and” is generally used between two target items.
Referential Decoupling: “2\n1.first glass is on the left\n2.second glass is in the middle\n3. third glass is on the right side”
Referent Grounding:

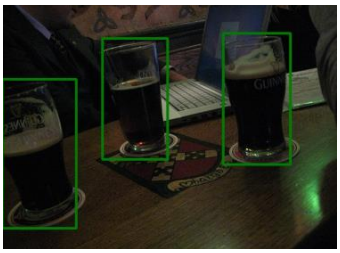


Table 9: Multi-target case.

Table 10: Unsupervised Schema. The bottom row is only using the generated data for training.

Training Set	gRefCOCO		
	val	testA	testB
gRefCOCO	39.61	32.19	36.44
Generated	32.29	25.85	27.91

Table 11: Ablation study on α.

α	RefCOCO			RefCOCO+			RefCOCOg	gRefCOCO		
	val	testA	testB	val	testA	testB	val	val	testA	testB
0.1	58.73	57.54	56.09	40.73	40.78	39.27	47.51	38.48	32.21	34.47
0.2	61.27	60.16	59.20	41.54	41.84	39.13	47.83	38.97	32.64	34.86
0.3	60.88	59.08	58.33	41.14	41.18	39.09	47.22	39.07	32.22	36.27
0.4	60.25	59.09	57.59	42.42	42.63	40.03	48.28	39.20	32.01	35.79
0.5	60.95	59.84	58.57	41.48	42.54	39.37	48.67	39.14	32.01	35.73
0.6	60.05	59.06	59.02	42.04	41.58	38.17	47.55	39.25	31.71	34.97
0.7	61.04	60.01	58.45	42.66	42.94	39.15	47.81	39.64	32.49	35.92
0.8	60.09	58.21	58.80	42.20	43.42	39.15	45.75	39.57	32.25	35.77
0.9	60.43	59.50	57.59	41.52	41.90	38.84	46.49	39.61	32.19	36.44

H More Quantitative Results

H.1 Unsupervised Generalized Referring Expression Comprehension

To demonstrate the robustness of our framework and explore its potential in a zero-annotation scenario, we extend LIHE to a fully unsupervised setting. In this setting, the model is trained without using any manually annotated data—neither annotated bounding boxes nor corresponding language queries. The only input is the image itself.

To achieve this, we first leverage the Vision-Language Model (VLM) to automatically generate a set of candidate "pseudo referring expressions" for each image. Subsequently, these machine-generated texts are used as the training data for the second stage of our model. As shown in the "Generated" row of Tab. 10, the experimental results demonstrate that despite relying entirely on machine-generated text, this unsupervised approach still achieves 32.29%/25.85%/27.91% performance on the gRefCOCO val/testA/testB splits, respectively. This result is only about 7-9% lower than its weakly supervised counterpart trained with authentic, human-annotated text, which strongly demonstrates LIHE's effectiveness in an annotation-free environment.

H.2 More Ablation Studies

Impact of the mixing weight α. Due to ρ being unknown, we conduct extensive experiments on the hybrid weight α as shown in Tab. 11 to find the best weight. Two clear trends emerge. *(i) U-shaped curve.* Extremely small α (closer to a pure Euclidean view) and extremely large α (approaching the hyperbolic-only view) both hurt performance; in every split, the scores first rise, peak in the mid-range, and then drop again. *(ii) Robust sweet-spot.* The interval $\alpha \approx 0.4$–0.7 consistently delivers the best or near-best numbers across all four benchmarks. For example, on RefCOCO testA the top accuracy of 60.01% is achieved at $\alpha = 0.7$, while gRefCOCO testB peaks at 36.44% with $\alpha = 0.9$, confirming that a balanced mixture captures complementary information from both geometries. Overall, the ablation validates the bias-variance analysis: an appropriate hybrid weighting outperforms either single-space similarity.

Explicit *vs.* Implicit Hierarchical Constraints. Tab. 13 investigates whether explicitly adding hierarchical losses helps. In the contrastive loss function (Eq. 3), similarity only involves the expression and the anchor vision feature, but we want to know whether it learns the whole hierarchy

Table 12: RefCLIP trained with/without $v = 0$ sample

RefCLIP trained	**WGREC**		
	val	testA	testB
With $v = 0$	16.77	15.78	20.53
Without $v = 0$	17.85	18.23	21.89

Table 13: Ablation on explicit hierarchical constraints on gRefCOCO.

Sim_E	Sim_H	Hierarchica Constraint	gRefCOCO		
			val	testA	testB
✓			38.88	31.77	34.89
	✓		39.58	31.57	36.88
✓	✓		39.61	32.70	35.84
✓		✓	39.08	32.33	35.17
	✓	✓	39.25	32.06	36.03
✓	✓	✓	39.37	32.44	36.62

implicitly. To validate this, we add an extra explicit hierarchical constraint loss as follows:

$$\mathcal{L}_{\text{hier}} = d_\kappa(f_{cat}, f_{base\_ref}) + d_\kappa(f_{ref}, f_{base\_ref})$$

where f_{cat} denotes the feature of the category text (e.g., 'person' as shown in 3), f_{base_ref} denotes the feature of the raw referring expression (e.g., 'left person') and f_{ref} denotes the feature of the decomposed referring expression. As shown in Tab. 13, adding the explicit hierarchical constraint basically has the same average performance, implying that through hyperbolic similarity, the model is able to learn Hierarchical Constraints Implicitly.

Training dataset with/without v=0 The Referent Grounding stage is based on the problem definition and optimization of RefClIP which do not support to handle v=0 samples: $a^* = \arg\max_{a \in \mathcal{A}} \phi(T, I, a)$. RefCLIP is based on the image-ref pair for contrastive loss, which the process of performing contrastive loss requires the inclusion of a positive sample; when v=0, there is no positive sample available for training the model, which will only damage the model performance. Therefore, in LIHE training stage, we directly remove the v=0 samples from Referent Grounding and in LIHE inference stage, v=0 samples are effectively addressed in the Referential Decoupling stage (Stage 1), where the model outputs "0" to indicate no detections.

I More Visualizations

In this section, we provide additional qualitative visualizations to further evaluate the behavior of the LIHE framework in various complex scenarios. As shown in Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, we demonstrate how the model successfully performs referential decoupling (RD) on the input referring expressions (RE) to accurately identify multi-target, single-target, or no-target situations. These diverse cases intuitively demonstrate the effectiveness and generalization capability of our proposed method.

RE: the right boy in black shirt is playing skateboard

RD: 0

RE: batter

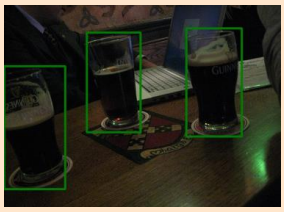
RD: batter: A child wearing a black helmet and white pants is swinging a bat at a baseball

RE: a guy in green and a rightmost guy

RD: 1.a guy in green 2.a rightmost guy

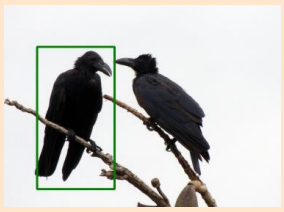
RE: three glasses

RD: 1.first glass is on the left side 2.second glass is in the middle 3. third glass is on the right side



RE: left bird

RD: Left bird: A black bird perched on the left side of the branch



RE: zebras

RD: 1.the zebra on the left side of the image 2.the zebra in the middle of the image 3.the zebra on the right side of the image

RE: both guys

RD: 1.the man brushing teeth 2.the man holding camera

RE: the guy wearing white coat on the leftmost with one hand holding the skateboard

RD: 0

RE: hand reaching in from left and guy

RD: 1.hand reaching in from left 2.guy

Figure 6: Qualitative visualizations of LIHE. RE denotes the original Referring Expression, and RD denotes the result of Referential Decoupling. The image is the result of Referent Grounding.

RE: sailor on phone on right, right man, and far left woman

RD: 1.the sailor on phone on right 2.the right man 3.the far left woman

RE: lady on the far right pretending to be interested

RD: Lady on the far right pretending to be interested

RE: the right boy in black shirt is playing skateboard

RD: 0

RE: goat on right

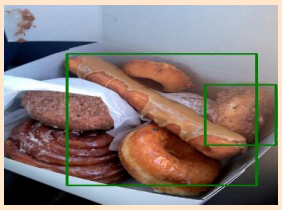
RD: 0

RE: whiterear end of horse

RD: A brown and white horse with a white patch on its ear is visible in the foreground

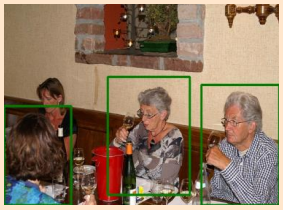
RE: long wiener thing and donut in the upper right corner of the box

RD: 1.long wiener thing 2.donut in the upper right corner of the box



RE: back of head on left, man on right, and middle woman

RD: 1.back of head on left 2.middle woman 3.man on right



RE: lady on left wearing visor

RD: lady on left wearing visor

RE: two men in black vests

RD: 1.the man on the left 2.the man second from the left

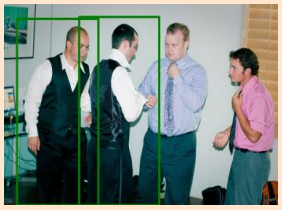


Figure 7: Qualitative visualizations of LIHE. RE denotes the original Referring Expression, and RD denotes the result of Referential Decoupling. The image is the result of Referent Grounding.

RE: guy in blue jacket/back to us

RD: 0

RE: the guy in the right wearing cap and glass with one knee on the ground

RD: 0

RE: the blue table holding four white plates

RD: 0

RE: standing

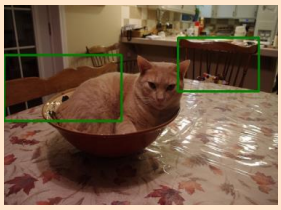
RD: The large elephant standing on the dirt ground

RE: two players

RD: 1.left person in red 2.red on right 3.shorts

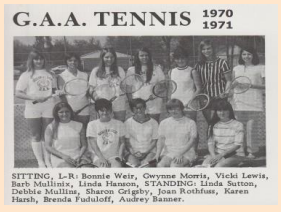
RE: chair at the end f table with purse n it and chair against table on left

RD: 1.chair at the end of the table with a purse on it 2.chair against the table on the left



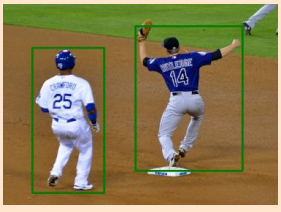
RE: the girl standing on the leftmost top row wearing white dress

RD: 0



RE: two players

RD: 1.player wearing blue jersey 2.player wearing white jersey



RE: man with arms up

RD: man with arms up

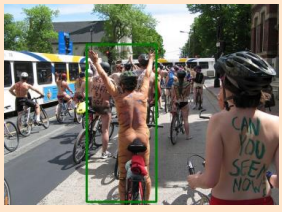


Figure 8: Qualitative visualizations of LIHE. RE denotes the original Referring Expression, and RD denotes the result of Referential Decoupling. The image is the result of Referent Grounding.

RE: two screens and the cat

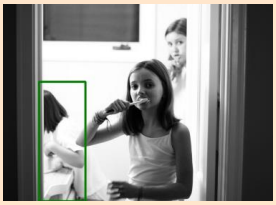
RD: 1.the left screen 2.the right screen 3.the cat

RE: the black laptop on the left

RD: 0

RE: Smaller child over sink

RD: Smaller child over sink

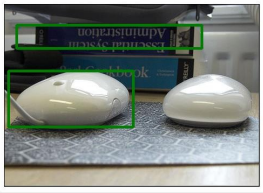


RE: the guy sitting at the back wearing glasses

RD: 0

RE: the mouse on the left and the book on the top

RD: 1.the mouse on the left 2.the book on the top



RE: a refree in black top is raising his arm out

RD: 0

RE: girl and MAN

RD: 1.the woman wearing yellow dress
2.the man in black suit

RE: man in blue shirt and tie

RD: Man wearing a light blue shirt and a patterned tie

RE: right front donut blonde

RD: 0

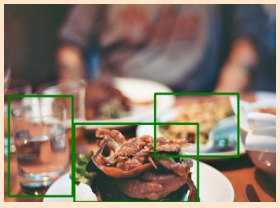
Figure 9: Qualitative visualizations of LIHE. RE denotes the original Referring Expression, and RD denotes the result of Referential Decoupling. The image is the result of Referent Grounding.

RE: a group of three individuals in the lower right corner

RD: 1.man in dark suit with arms crossed 2.woman in gray dress holding wine glass 3.man in pinstripe suit facing away

RE: blurry food behindslight right of dish in forefront, food front middle, and the glass on left next to plate

RD: 1.blurry food behindslight right of dish in forefront 2.food front middle 3.the glass on left next to plate



RE: white shirt woman in front of group left, man pushing cart, and man on left with long pants

RD: 1.white shirt woman in front of group left 2.man pushing cart 3.man on left with long pants

Figure 10: Qualitative visualizations of LIHE. RE denotes the original Referring Expression, and RD denotes the result of Referential Decoupling. The image is the result of Referent Grounding.