A-PAIR: A BENCHMARK AND IDENTITY-CONSISTENT GROUNDING FRAMEWORK FOR AIR-GROUND CROSS-VIEW REFERRING PERSON DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Air-ground cross-view referring person detection is a necessary component in the language-to-perception-to-control chain of collective embodied intelligence, grounding a language command into the same physical target before ground and aerial agents can coordinate downstream actions. Existing referring expression comprehension and open-vocabulary grounding methods do not jointly account for cross-view identity consistency, making them insufficient for Air-Ground Cross-View Referring Person Detection (AGCV-RPD), which involves similar pedestrian distractors, weak aerial appearance cues, and cross-view identity consistency. To study this problem, we introduce Air-Ground Paired Identity-Aware Referring (A-PAIR), a benchmark for full-image air-ground cross-view referring person detection, containing 22,137 cross-view referring samples. To construct A-PAIR efficiently, we propose Factorized Annotation and Referential Alignment (FARA), a semi-automatic annotation framework that generates factorized referring descriptions and identity-consistency supervision at reduced cost. We propose Identity-Consistent Referring Grounding (ICRG), a framework that combines factorized referential grounding, candidate-completeness supervision, and cross-view consistency calibration for joint air-ground pair selection. ICRG improves ground, aerial, and pair-level detection over strong baselines, increasing pair F1 from 16.65% to 22.28%. AGCV-RPD requires paired detection and identity-consistent reasoning.

INTRODUCTION

Air-ground cross-view referring person detection is a necessary component in the language-to-perception-to-control chain of collective embodied intelligence: a language command must first be grounded to the same physical target across complementary sensors before ground and aerial agents can coordinate downstream actions. For such collaborative embodied systems, a fundamental capability is to understand a natural-language referring expression and ground it consistently across heterogeneous sensors. A human or embodied agent may describe a person once, yet expect both the ground and aerial views to detect the same physical individual. We refer to this task as Air-Ground Cross-View Referring Person Detection (AGCV-RPD). Formally, given a ground image, an aerial image of the same scene, and a referring expression, the model must output one bounding box per view that detects the same person identity, as illustrated in Fig. 1. Solving AGCV-RPD links free-form language, cross-view perception, and identity reasoning into a foundational query interface for air-ground embodied collaboration.

Existing referring expression comprehension and open-vocabulary grounding methods, including TransVG Deng et al. (2021), MDETR Kamath et al. (2021), GLIP Li et al. (2022), GroundingDINO Liu et al. (2024), and RefDrone Sun et al. (2026), provide useful priors for single-view detection. However, AGCV-RPD is not a direct extension of single-image grounding. Similar pedestrians and repeated spatial layouts create target ambiguity, while small targets in aerial views often lack sufficient visual cues. More importantly, two individually plausible predictions may correspond to different people, whereas standalone identity matching assumes that candidate boxes have already been generated. Air-ground referring is therefore fundamentally a coupled detection and identity-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

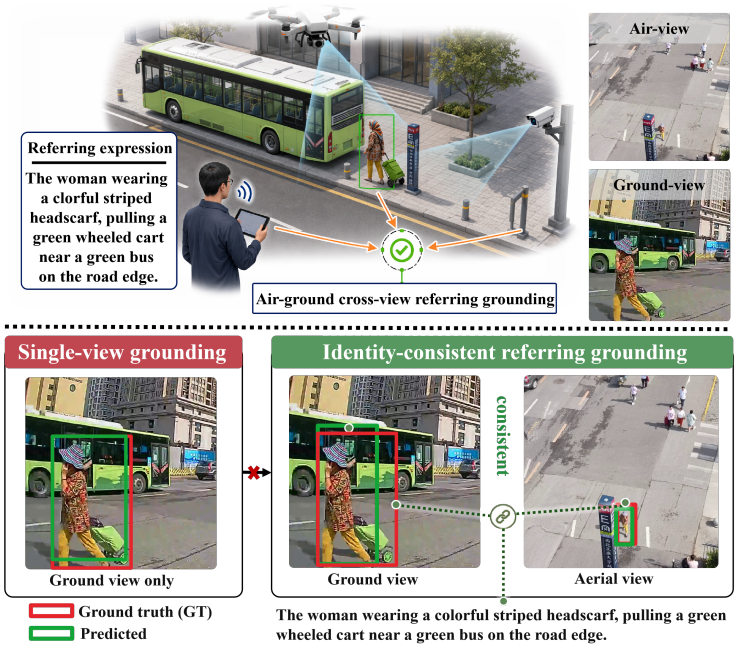


Figure 1: AGCV-RPD task. The model detects the target person in ground and aerial views.

consistency problem rather than two independent single-view grounding tasks. This gives rise to three key challenges:

1. Similar pedestrians and spatial ambiguity.
2. Weak appearance cues of small targets.
3. Cross-view identity consistency.

We address these challenges with Identity-Consistent Referring Grounding (ICRG), an identity-consistent referring grounding framework. Rather than appending a standalone retrieval module after grounding, ICRG builds a unified query-to-pair prediction pipeline around three complementary supervision signals. Factorized referential supervision separates identity-stable appearance semantics from view-dependent scene-spatial semantics to reduce ambiguity among similar pedestrians. Candidate-completeness supervision expands the candidate space so small or weakly described targets are not removed before pair reasoning. Auxiliary cross-view identity-consistency supervision calibrates candidate pairs and suppresses boxes that are individually plausible but jointly identity-inconsistent. To enable this study, we propose Factorized Annotation and Referential Alignment (FARA) to construct Air-Ground Paired Identity-Aware Referring (A-PAIR), a benchmark derived from the G2APS air-ground person dataset with paired images, same-identity boxes, structured text, and auxiliary identity-consistency labels.

Our contributions are as follows:

- A-PAIR benchmark. We introduce A-PAIR, a full-image benchmark for AGCV-RPD. A-PAIR contains 22,137 cross-view referring samples over 7,588 unique ground images and 3,891 unique aerial images.
- FARA annotation framework. We propose FARA, a semi-automatic pipeline combining VLM annotation, keyframe propagation, and quality control for cross-view descriptions and identity supervision.
- ICRG framework. We propose ICRG for the AGCV-RPD task. ICRG integrates candidate-completeness supervision, factorized referential grounding, and cross-view consistency calibration to improve cross-view matching performance in complex air-ground scenes.

RELATED WORK

LANGUAGE-GUIDED PERSON DETECTION

Language-guided person detection localizes a person described by natural language. Existing methods mainly include **(1) context-aware grounding**, which exploits attributes, spatial context, and inter-object relations Yu et al. (2018); Mao et al. (2016); Hu et al. (2016); **(2) Transformer-based direct grounding**, which aligns text and image regions through cross-modal attention, box regression, or set prediction Deng et al. (2021); Kamath et al. (2021); Deng et al. (2023); and **(3) region-language pre-training and person-centric grounding**, which leverages large-scale vision-language alignment, person search, and small-object modeling Li et al. (2022); Liu et al. (2024); Dai et al. (2025); Li et al. (2017); Zhang et al. (2023a); Goto et al. (2025); Zhan et al. (2026).

However, most studies localize a referred person within a single image without enforcing cross-view identity consistency. AGCV-RPD instead formulates paired language-guided detection, while A-PAIR provides same-identity ground–aerial annotations and pair-level evaluation.

CROSS-VIEW AIR-GROUND PERCEPTION

Cross-view air-ground perception establishes correspondence across substantially different viewpoints. Existing approaches include **(1) scene-level geo-localization**, which matches ground images with aerial or satellite references Workman et al. (2015); Hu et al. (2018); Zhu et al. (2021); **(2) identity-level cross-view matching**, including person re-identification and air-ground person search Luo et al. (2019); Sun et al. (2018); He et al. (2021); Ye et al. (2022); Zhang et al. (2023b); Zhou et al. (2026); and **(3) multi-view language grounding**, covering multi-image grounding, multi-perspective communication, cross-view referring comprehension, and aerial referring detection Xiao et al. (2025); Li et al. (2025); Tang et al. (2024); Li et al. (2026); Sun et al. (2026).

These methods address place matching, identity retrieval, or language grounding, but rarely jointly localize the same referred person in paired full-image ground and aerial views. A-PAIR fills this gap with paired bounding boxes and cross-view identity links, while ICRG combines factorized grounding, candidate-completeness supervision, and identity-consistency calibration. Table 2 summarizes the differences from related benchmarks..

THE A-PAIR BENCHMARK

DATA SOURCE

A-PAIR is derived from G2APS Zhang et al. (2023b). From paired frames, we retain cross-view observations that share at least one person identity, yielding $7{,}588$ shared-identity ground-aerial pairs associated with $7{,}588$ unique ground images and $3{,}891$ unique aerial images. All bounding boxes follow the $xywh$ convention. Each sample contains both an appearance-full expression and a scene-spatial expression, enabling full-, spatial-, and factorized-prompt evaluation within a unified protocol. To further ensure annotation quality, we manually reviewed all annotations for visual consistency, target alignment, and referential uniqueness.

FARA ANNOTATION

FARA is a semi-automatic pipeline that constructs A-PAIR from paired G2APS observations, grouping frames by scene and track and pairing temporally aligned ground and aerial views with their shared-identity boxes. A VLM generates a reusable appearance phrase for each identity and view-specific spatial phrases for each view; together, these form the referring expressions. FARA also exports all-person detection labels and cross-view identity pairs to supervise candidate completeness and identity consistency. To prevent image overlap across partitions, samples sharing an image are assigned to the same split through connected-component grouping.

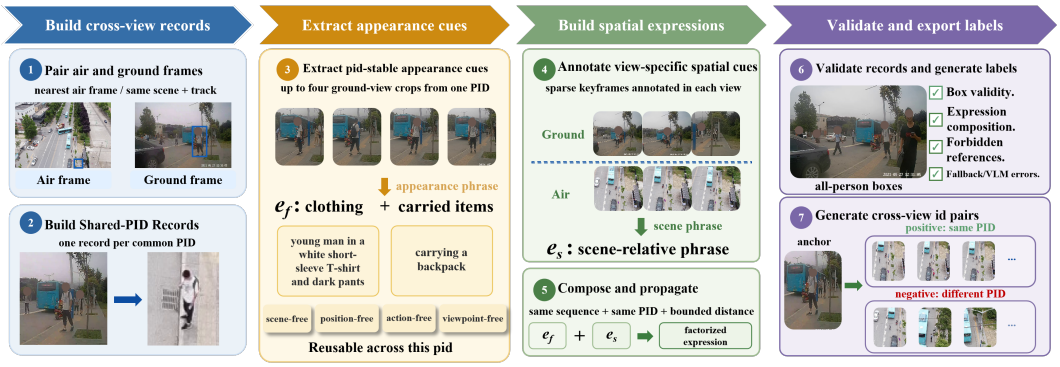


Figure 2: FARA annotation pipeline. The pipeline constructs A-PAIR from paired G2APS observations with factorized descriptions and identity labels.

Split	Samples	Pairs	Gnd. img.	Aer. img.
Train	15,497	6,672	6,672	3,530
Val	2,213	279	279	69
Test	4,427	637	637	292
Total	22,137	7,588	7,588	3,891

Table 1: A-PAIR split statistics.

Data Statistics

Table 1 summarizes the split of A-PAIR. It contains 22,137 cross-view referring samples, where each sample is a referring instance with paired ground–aerial images and a target identity; 15,497 are used for training, 2,213 for validation, and 4,427 for testing. The PIDs column counts unique cross-view identities, the Pairs column counts unique ground–aerial image pairs, and the Gnd. img. and Aer. img. columns count unique ground and aerial images, respectively.

Benchmark Comparison

As summarized in Table 2, existing benchmarks cover different subsets of the capabilities required by AGCV-RPD. RefCOCO and RefDrone support language conditioned localization, while MIG-Bench extends grounding to multiple images. CVUSA and VIGOR provide paired ground–aerial observations for cross-view geo-localization, and G2APS provides air-ground identity-linked data without language input. TAG-PEDES further introduces text-based retrieval across heterogeneous aerial and ground views, but its primary output is a ranked person image rather than bounding boxes in paired full images. In contrast, A-PAIR combines language input, paired ground–aerial images, full-image localization, same-identity box annotations, cross-view identity links, and pair-level evaluation in one benchmark.

Evaluation Metrics

Using the hit definition from the Task Definition subsection, we report instance-level and image-level F1 and accuracy for each view. In this single-target referring protocol, every other person in the image counts as a negative. The instance-level metrics are defined as

$$\mathrm{F1}_{\mathrm{inst}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{Acc}_{\mathrm{inst}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{2}$$

where $\mathrm{TN} = 0$ because there are no no-target samples. Image-level metrics ($\mathrm{F1}_{\mathrm{img}}$, $\mathrm{Acc}_{\mathrm{img}}$) score whether the top prediction attains

$$\mathrm{IoU}(\hat{b}_v, b_v^*) \geq 0.5. \tag{3}$$

Dataset	View(s)	Lang.	Lang.-Box Loc.	ID Supervision	Pair Eval.
RefCOCO Yu et al. (2016)	Ground	✓	✓	-	-
RefCOCO+ Yu et al. (2016)	Ground	✓	✓	-	-
RefCOCOg Mao et al. (2016)	Ground	✓	✓	-	-
RefDrone Sun et al. (2026)	Aerial	✓	✓	-	-
VisDrone Du et al. (2019)	Aerial	-	-	-	-
UAVDT Du et al. (2018)	Aerial	-	-	-	-
DOTA Xia et al. (2018)	Aerial	-	-	-	-
CVUSA Workman et al. (2015)	Ground+Aerial	-	-	-	-
VIGOR Zhu et al. (2021)	Ground+Aerial	-	-	-	-
Market-1501 Zheng et al. (2015)	Ground	-	-	ID labels	-
DukeMTMC Ristani et al. (2016)	Ground	-	-	ID labels	-
G2APS Zhang et al. (2023b)	Ground+Aerial	-	-	Same-ID boxes	-
A-PAIR	Ground+Aerial	✓	✓	Same-ID boxes	✓

Table 2: Benchmark comparison. Lang.-Box Loc. denotes language-conditioned localization in an uncropped image. ID Supervision describes identity-level annotations, including identity labels and same-identity bounding boxes across views. Pair Eval. denotes joint evaluation of view-specific predictions as a paired instance.

For pair-level evaluation, a prediction is correct only when both views are correct:

$$\mathrm{hit}_{\text{pair}} = \mathrm{hit}_g \wedge \mathrm{hit}_a. \tag{4}$$

The pair metrics (Pair Acc, Pair F1) are therefore the primary indicator of cross-view competence.

METHOD

ICRG addresses AGCV-RPD with a unified framework for language guided pair detection, as illustrated in Fig. 3. Given paired ground and aerial images and factorized referring expressions, it first defines the pair correctness criterion, then builds candidate pools through factorized grounding and all person completion, and finally uses auxiliary identity consistency to calibrate candidate pairs before selecting the final ground and aerial boxes. The following subsections formalize the task, describe the three scoring sources, and present the fusion inference procedure.

TASK DEFINITION

Each A-PAIR sample is a tuple (I_g, I_a, e_f, e_s, y), where I_g and I_a are paired ground and aerial images of a shared scene, e_f is an appearance-full referring expression, e_s is a scene-spatial expression, and y is the target person identity (pid). The model must predict one box per view, $(\hat{b}_g, \hat{b}_a)$, both detecting the person with identity y. A single-view prediction is correct when its intersection-over-union with the ground-truth box exceeds a threshold $\tau = 0.5$:

$$\mathrm{hit}_v = \mathbb{1}\Big[\mathrm{IoU}(\hat{b}_v, b_v^*) \geq \tau\Big], \quad v \in \{g, a\}. \tag{5}$$

A pair prediction is correct only when both views are correct, $\mathrm{hit}_g \wedge \mathrm{hit}_a$, which makes the pair criterion harder than either view. This coupling is what single-view transfer fails to satisfy and motivates the design below.

FACTORIZED GROUNDING

The first supervision signal detects candidates referred to by language in each view. Rather than treating a referring expression as an indivisible sentence, A-PAIR factorizes the supervision into an appearance profile for each identity and a scene and spatial description for each view. The full appearance expression e_f captures stable identity semantics, whereas the scene and spatial expression e_s retains spatial cues relative to scene landmarks, such as being near a bus stop shelter, that remain useful when appearance is weak. We feed both factors to a grounding model G built on the GroundingDINO architecture Liu et al. (2024) with a Swin Transformer image encoder and a text encoder

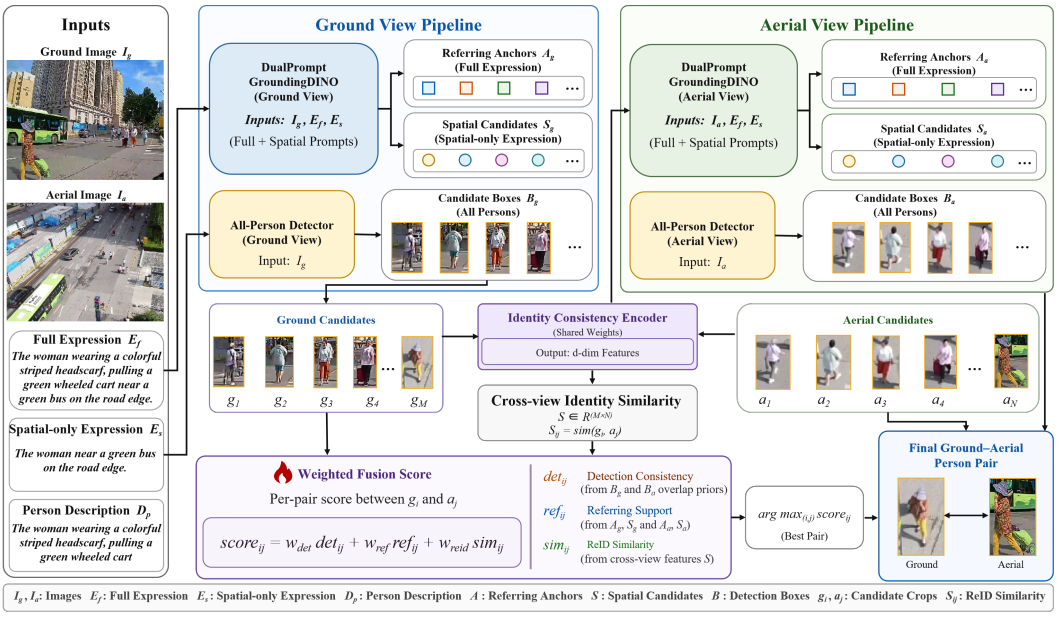


Figure 3: ICRG framework. It combines factorized grounding, candidate completion, and cross-view consistency.

coupled by cross modal attention. For each view v and factor $e \in \{e_f, e_s\}$, the model produces scored referring boxes

$$\mathcal{R}_v^e = G(I_v, e) = \{(r_k, s_k^{\text{ref}})\}_{k=1}^{K}, \tag{6}$$

where r_k is a candidate box and s_k^{ref} its referring confidence. The grounding model is trained with the standard grounded detection objective, which combines a contrastive token alignment loss and box regression:

$$\mathcal{L}_{\text{ground}} = \mathcal{L}_{\text{align}} + \lambda_{L1}\,\mathcal{L}_{L1} + \lambda_{giou}\,\mathcal{L}_{giou}. \tag{7}$$

The appearance factor drives precise detection when appearance is visible, while the scene and spatial factor improves recall in crowded scenes where appearance is ambiguous, providing complementary candidates $\mathcal{R}_v = \mathcal{R}_v^{e_f} \cup \mathcal{R}_v^{e_s}$. This design uses appearance to constrain who the referred person could be and spatial context to constrain where that person can plausibly appear in each view. Taking the union rather than forcing a single expression avoids committing to one cue prematurely and defers the final identity decision to the fusion stage, where evidence from both views is available.

Candidate Completeness

Referring grounding can still miss the target when the expression is under-specified or the target is small. To bound this failure, candidate-completeness supervision trains an all-person detector D, sharing the same detector backbone but optimized on the A-PAIR all-person detection labels, to enumerate every person in each view:

$$\mathcal{P}_v = D(I_v) = \{(p_j, s_j^{\text{det}})\}_{j=1}^{M_v}, \tag{8}$$

with detection confidence s_j^{det}. D is optimized with the detection loss $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}}$. The union $\mathcal{C}_v = \mathcal{R}_v \cup \mathcal{P}_v$ forms the candidate pool per view, so a correct target that the referring signal overlooks can still enter the final selection. The role of D is recall-oriented rather than precision-oriented: because pair selection is decided after cross-view calibration, the detector only needs to place the target somewhere in $\mathcal{C}_v$, and any spurious boxes it introduces are filtered later by referring and consistency scores.

Cross-View Consistency

The third supervision signal calibrates whether two language-compatible candidates also refer to the same physical person. For a candidate box c in view v, we crop the region and embed it with a

Algorithm 1 ICRG inference. Candidate pairs are scored by fused detection, referring, and identity-consistency cues.

Require: ground image I_g, aerial image I_a, prompts e_f, e_s
1: $\mathcal{R}_v \leftarrow G(I_v, e_f) \cup G(I_v, e_s)$ for $v \in \{g, a\}$
2: $\mathcal{P}_v \leftarrow D(I_v)$; $\mathcal{C}_v \leftarrow \mathcal{R}_v \cup \mathcal{P}_v$
3: **for all** $(g_i, a_j) \in \mathcal{C}_g \times \mathcal{C}_a$ **do**
4: compute $s_{ij}^{\mathrm{det}}, s_{ij}^{\mathrm{ref}}, s^{\mathrm{id}}(g_i, a_j)$
5: $\mathrm{score}(g_i, a_j) \leftarrow$ Eq. (11) with val-selected weights
6: **end for**
7: **return** $\arg\max_{(g_i,a_j)} \mathrm{score}(g_i, a_j)$

ResNet-50 encoder f into a unit-normalized feature $f(c) \in \mathbb{R}^d$. The auxiliary identity-consistency score between a ground candidate g_i and an aerial candidate a_j is the cosine similarity

$$s^{\mathrm{id}}(g_i, a_j) = \frac{f(g_i)^\top f(a_j)}{\|f(g_i)\| \, \|f(a_j)\|}. \tag{9}$$

The encoder is trained on cross-view crop pairs from the training split using the binary consistency loss

$$\begin{aligned} \mathcal{L}_{\mathrm{id}} = -\sum_{(i,j)\in\mathcal{T}} \Big[& y_{ij} \log \sigma\big(s_{ij}^{\mathrm{id}}\big) \\ & + (1 - y_{ij}) \log\big(1 - \sigma\big(s_{ij}^{\mathrm{id}}\big)\big) \Big], \end{aligned} \tag{10}$$

where $\mathcal{T}$ is the set of training cross-view crop pairs, $y_{ij} = \mathbb{1}[\mathrm{pid}(g_i) = \mathrm{pid}(a_j)]$ indicates whether the two crops share the same pid, and σ is the logistic function. The signal supplies cross-view identity consistency that language alone cannot provide. It learns to bridge the appearance gap between a street-level crop, where clothing and build are visible, and an aerial crop of the same person, where those attributes are lost to scale and viewpoint. The signal is auxiliary: it does not retrieve a person independently of language, but calibrates pairs proposed by the referring and completeness signals. It can suppress an identity-inconsistent pair in which each view matches the expression but both boxes depict different people.

Fusion and Inference

The final stage selects one identity-consistent ground-aerial candidate pair. For each candidate pair $(g_i, a_j) \in \mathcal{C}_g \times \mathcal{C}_a$ we compute a fused score

$$\mathrm{score}(g_i, a_j) = w_{\mathrm{det}} \, s_{ij}^{\mathrm{det}} + w_{\mathrm{ref}} \, s_{ij}^{\mathrm{ref}} + w_{\mathrm{id}} \, s^{\mathrm{id}}(g_i, a_j), \tag{11}$$

where s_{ij}^{det} and s_{ij}^{ref} aggregate the detection and referring confidences of the two boxes. The non-negative weights $(w_{\mathrm{det}}, w_{\mathrm{ref}}, w_{\mathrm{id}})$ are not learned by gradient descent; they are selected by grid search on the validation split to maximize pair accuracy, which keeps the fusion interpretable and avoids overfitting the small test set. At inference, the predicted pair is the arg-max of Eq. (11):

$$(\hat{b}_g, \hat{b}_a) = \arg\max_{(g_i,a_j)} \mathrm{score}(g_i, a_j). \tag{12}$$

Algorithm 1 summarizes inference. All learnable components (G, D, f) are trained on the training split with their objectives; the fusion weights are fixed once selected on validation, and the test split is evaluated once.

Experiments

Setup

Dataset, protocol, and metrics. All experiments use the A-PAIR dataset and the evaluation protocol fromTable 2 summarizes the differences from related benchmarks. Evaluation Metrics, with $15{,}497$ training, $2{,}213$ validation, and $4{,}427$ test samples. Unless noted, methods are trained on the training split and evaluated once on the test split. We report the per-view instance-level and image-level metrics and the pair-level metrics defined in Evaluation Metrics.

Method	Ground				Aerial				Pair	
	$F1_i$	Acc_i	$F1_{img}$	Acc_{img}	$F1_i$	Acc_i	$F1_{img}$	Acc_{img}	Acc	F1
TransVG Deng et al. (2021)	2.53	1.28	4.94	2.53	0.05	0.02	0.09	0.05	0.00	0.00
PropVG Dai et al. (2025)	1.49	0.75	2.94	1.49	0.00	0.00	0.00	0.00	0.00	0.00
Vision-R1 Zhan et al. (2026)	1.92	0.97	3.77	1.92	0.00	0.00	0.00	0.00	0.00	0.00
SOREC-PIZA Goto et al. (2025)	13.96	7.50	24.50	13.96	7.43	3.86	13.84	7.43	1.78	3.51
GDINO-B Liu et al. (2024)	26.74	15.44	42.20	26.74	12.02	6.39	21.46	12.02	5.74	10.85
RefDrone Sun et al. (2026)	*33.50*	*20.12*	*50.19*	*33.50*	14.66	7.91	25.57	14.66	8.31	15.35
GDINO-T Liu et al. (2024)	32.28	19.25	48.80	32.28	*17.42*	*9.54*	*29.67*	*17.42*	*9.08*	*16.65*
ICRG (ours)	**35.17**	**21.34**	**52.04**	**35.17**	**21.89**	**12.29**	**35.92**	**21.89**	**12.54**	**22.28**

Table 3: Main results on A-PAIR. All metrics are percentages; bold and italics mark the best and second-best results.

Fact.	Cand.	Cons.	Ground		Aerial		Pair	
			$F1_i$	$F1_{img}$	$F1_i$	$F1_{img}$	Acc	F1
-	-	-	7.32	13.65	19.91	33.21	1.83	3.60
✓	-	-	34.88	51.72	20.83	34.47	11.66	20.88
✓	✓	-	**36.12**	**53.07**	21.82	35.82	12.24	21.82
✓	✓	✓	35.17	52.04	**21.89**	**35.92**	**12.54**	**22.28**

Table 4: Ablation results on A-PAIR. Columns indicate enabled signals under the same protocol.

IoU-Quality Diagnosis Reveals Core Challenges

Figures 4 diagnose the main failure modes of AGCV-RPD. A sample is counted as a successful pair-level detection only when both the ground and aerial predictions satisfy $\mathrm{IoU} \geq 0.5$; this occurs for only 12.54% of the test set, showing that cross-view pair detection is substantially harder than single-view detection. The IoU decomposition shows that some failures fall into partial-overlap or near-target bins, while the dominant errors are missed aerial targets and ground-correct/aerial-missed cases. These observations motivate explicit modeling of cross-view identity consistency in AGCV-RPD.

Main Results Validate Paired Reasoning

Table 3 compares ICRG with seven single view baselines on the image disjoint test split, including REC, small object grounding, and VLM-based detection models. The strongest baseline, GroundingDINO-T Liu et al. (2024), obtains 16.65% pair F1, while ICRG improves it to 22.28%. This gain shows that independently grounding the two views is insufficient for AGCV-RPD, because two plausible single view predictions may still fail to form an identity consistent pair. The weak pair results of recent methods such as SOREC-PIZA, PropVG, and Vision-R1 further support this conclusion. As visualized in Fig. 4, ICRG achieves the best tradeoff between ground view detection and pair level performance, validating the need for joint language grounding, candidate recall, and cross view consistency.

Factorized Grounding Reduces Ambiguity

Factorized grounding addresses similar pedestrians and spatial ambiguity. In Table 4, enabling only Fact. improves pair F1 from 3.60% to 20.88%. This indicates that separating appearance and spatial cues is more effective than treating the referring expression as an undifferentiated prompt. The appearance factor narrows who the target could be, while the scene and spatial factor narrows where the target can plausibly appear in each view. Their union provides complementary candidates and reduces mistakes caused by ambiguous descriptions or visually similar pedestrians.

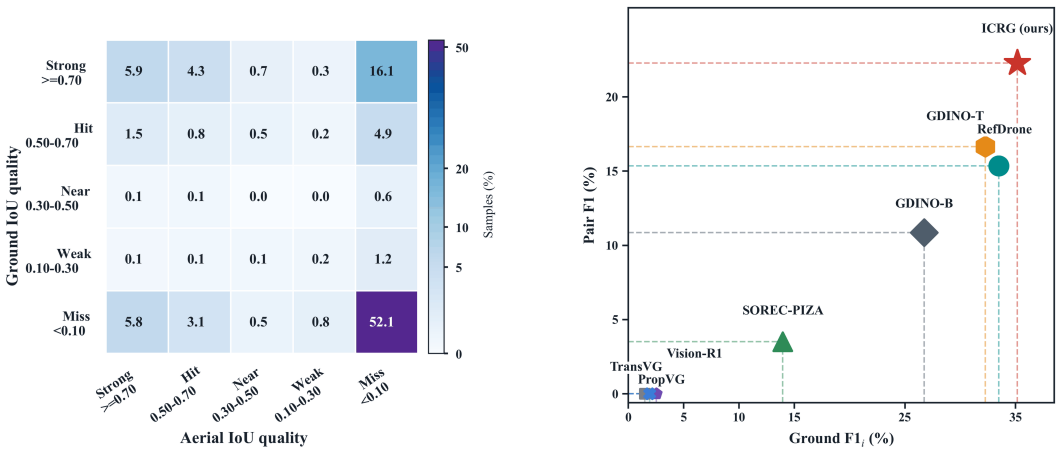


Figure 4: Left: pair-level IoU-quality decomposition. Each cell reports the percentage of A-PAIR test samples in a ground-aerial IoU-quality bin. Right: ground versus pair performance. Each point shows one method by Ground F1$_i$ and Pair F1 on A-PAIR.

Completeness Improves Aerial Recall

Candidate completeness mainly benefits aerial detection, where targets are small and appearance cues are weak. In Table 4, adding Cand. to the setting with Fact. alone raises aerial F1$_i$ from 20.83% to 21.82% and aerial F1$_{\text{img}}$ from 34.47% to 35.82%, with the full model further reaching 21.89% and 35.92%. The larger image-level gain suggests that Cand. recovers valid targets in difficult aerial frames, while the smaller instance-level change indicates that additional proposals may also introduce distractors. The all-person detector therefore serves as a recall safeguard. When the referring branch misses a tiny or under-specified aerial target, candidate completion can still retain that person for later pair selection. By decoupling proposal recall from language confidence, Cand. prevents an early grounding miss from becoming irreversible. Its gain thus comes from improving target availability rather than acting as a standalone precision module.

Consistency Calibration Improves Pairing

Consistency calibration improves the final pair decision after candidate generation produces plausible boxes in both views. In Table 4, adding Cons. to the Fact.+Cand. setting raises pair Acc from 12.24% to 12.54% and pair F1 from 21.82% to 22.28%. These gains are modest but meaningful because Cons. does not create new candidates; it reranks the existing ground–aerial candidate pairs according to identity compatibility. Thus, identity consistency changes the final decision from selecting two separately plausible boxes to selecting a jointly valid pair, although missed aerial targets and similar distractors still limit absolute pair accuracy.

Conclusion

We introduced air-ground cross-view referring person detection and developed A-PAIR, a benchmark built from paired ground-aerial observations with same-identity boxes, factorized language expressions, all-person detection labels, and identity-consistency pairs. A-PAIR formulates language-guided person detection as a joint cross-view decision problem: the predicted ground and aerial boxes must both match the query and correspond to the same physical individual. Existing single-view referring methods process the two views independently, making them vulnerable to ambiguous candidates and cross-view identity mismatches. To address this challenge, ICRG unifies factorized referential grounding, candidate-completeness supervision, and cross-view consistency calibration in a query-to-pair framework. On the A-PAIR dataset, ICRG consistently improves ground-view, aerial-view, and pair-level detection over strong single-view baselines, increasing pair F1 from 16.65% to 22.28%. These results demonstrate the value of explicitly modeling cross-view iden-

tity consistency for paired referring detection. Future work will extend A-PAIR toward temporal air-ground observations, more crowded scenes, and broader environmental diversity.

REFERENCES

Ming Dai, Wenxuan Cheng, Jiedong Zhuang, Jiang-Jiang Liu, Hongshen Zhao, Zhenhua Feng, and Wankou Yang. PropVG: End-to-end proposal-driven visual grounding with multi-granularity discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7058–7068, 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. 1749–1759, 2021.

Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yang 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.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision*, pp. 370–386, 2018.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 213–226, 2019.

Kanoko Goto, Takumi Hirose, Mahiro Ukai, Shuhei Kurita, and Nakamasa Inoue. Referring expression comprehension for small objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21231–21242, 2025.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14993–15002, 2021.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555–4564, 2016.

Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267, 2018.

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. 1760–1770, 2021.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10955–10965, 2022.

Liuwu Li, Ronger Ding, Zehang Lin, Qi Peng, and Jiayuan Xie. Subject-agnostic cross-view referring expression comprehension via reward-driven consistency learning. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1324–1332, 2026.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5187–5196, 2017.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, and Maosong Sun. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9845–9867. Association for Computational Linguistics, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55, 2024.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1487–1495, 2019.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan 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.

Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 17–35, 2016.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling and a strong convolutional baseline. In *Proceedings of the European Conference on Computer Vision*, pp. 480–496, 2018.

Zhichao Sun, Yepeng Liu, Huachao Zhu, Yuliang Gu, Yuda Zou, Zelong Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. RefDrone: A challenging benchmark for referring expression comprehension in drone scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1752–1762, 2026.

Zineng Tang, Lingjun Mao, and Alane Suhr. Grounding language in multi-perspective referential communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19727–19741, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3961–3969, 2015.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3): 2749–2771, 2025.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*, pp. 69–85, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-R1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5807–5817, 2026.

Shizhou Zhang, De Cheng, Wenlong Luo, Yinghui Xing, Duo Long, Hao Li, Kai Niu, Guoqiang Liang, and Yanning Zhang. Text-based person search in full images via semantic-driven proposal generation. In *Proceedings of the 4th International Workshop on Human-Centric Multimedia Analysis*, pp. 5–14, 2023a.

Shizhou Zhang, Qingchun Yang, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Ground-to-aerial person search: Benchmark dataset and approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 789–799, 2023b.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116–1124, 2015.

Xinyu Zhou, Yu Wu, Jiayao Ma, Wenhao Wang, Min Cao, and Mang Ye. Text-based aerial-ground person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29035–29043, 2026.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5316–5325, 2021.