SARA: Semantic-Anchored Referential Alignment for Ego–Exo Instance Correspondence

Yueyang Ge, **Ye Lin**, **Bojun Yang**, **Yilin Huang** and **Yi Guo***
School of Automation and Intelligent Sensing, Shanghai Jiao Tong University
{yueyangge, ye.lin, yangbojun, hyl1101, guo.yi}@sjtu.edu.cn

Abstract

Achieving visual coordination between egocentric (ego) and exocentric (exo) perspectives is a cornerstone of augmented reality and human-machine collaboration. However, extreme viewpoint shifts and occlusions often cause semantic drift in existing methods, making robust cross-view correspondence a formidable challenge. In this paper, we propose SARA (Semantic-Anchored Referential Alignment), a framework designed to bridge the ego-exo gap by leveraging viewpoint-invariant, high-level semantic information as region-level semantic anchors for robust feature mapping. Specifically, our approach incorporates: (1) the Semantic Anchoring Module extracts cross-view category priors and invariant features to provide stable region-level references across perspectives, and (2) the Multimodal Semantic Vector Fusion mechanism achieves language-guided manifold fusion of semantic vectors to synthesize unified embeddings, effectively establishing representational correspondence across disparate perspectives. Extensive experiments across diverse, complex scenarios demonstrate SARA's superior performance in establishing stable mappings, validating the effectiveness of semantic-invariant features and multimodal fusion in addressing the core hurdles of cross-view understanding.

1 Introduction

In augmented reality (AR) and human–machine collaboration [Robinson *et al.*, 2023], agents often need to associate objects across egocentric (ego) and exocentric (exo) views for shared perception and assistance [Sun *et al.*, 2025]. A typical scenario is instruction-driven grounding: a user specifies a target in the ego view, and the system must localize and link the same physical instance in the exo view. Ego–exo instance correspondence therefore requires identity-preserving association across viewpoints while reflecting user intent.

Despite its importance, ego–exo correspondence is brittle in unconstrained scenes. Drastic viewpoint and scale changes, together with occlusion and clutter, often make local appearance cues unreliable, leading to look-alike confusion and semantic drift where the target is matched to an identity-mismatched instance [Zhang *et al.*, 2022; Liu *et al.*, 2025]. Moreover, partial visibility or small targets further weaken evidence for precise association, making cross-view decisions hard to audit and unstable under distribution shifts.

Existing methods either localize the target using segmentation/open-vocabulary models or rank candidate masks via learned cross-view representations [Zhang *et al.*, 2025]. However, viewpoint changes and occlusion still limit generalization and robust reasoning.

To address these issues, this paper proposes SARA (Semantic-Anchored Referential Alignment) for ego–exo instance correspondence. SARA combines a Semantic Anchoring Module and Language-conditioned Multimodal Fusion to tackle both appearance ambiguity and semantic drift. The anchoring module builds a high-recall candidate pool and discovers shared reference anchors to provide spatial structure, while the fusion module integrates visual features, instance–anchor geometry, and instruction cues for robust matching under viewpoint changes. Final predictions are made via candidate retrieval with a reject option for unreliable matches.

SARA is evaluated under challenging conditions with dense similar instances, small targets, and heavy occlusion. It consistently outperforms strong baselines in matching accuracy and stability under challenging conditions.

Contributions are summarized as follows:

- **SARA is proposed** as a semantic-anchored referential alignment framework for ego–exo instance correspondence, upgrading appearance matching to an interpretable reasoning-based retrieval formulation.
- **Semantic anchoring with structural reasoning** is developed by combining ROI-guided candidate mining and cross-view reference anchor discovery with topology encoding and language-conditioned fusion, enforcing both semantic and structural consistency.
- **Robust generalization is validated** on Ego-Exo4D and challenging settings, demonstrating consistent gains in accuracy and stability.

*Corresponding author.

2 Related Works

Cross-view Instance Correspondence. Cross-view instance correspondence targets identity-preserving matching of the *same instance* under large viewpoint changes, enabling target transfer across views for localization and segmentation. With time-synchronized first-/third-person datasets (e.g., Ego-Exo4D [Grauman *et al.*, 2024]), this challenge is typically framed as predicting the exocentric instance mask based on an egocentric target-mask query, where preserving identity takes precedence over maintaining category consistency. Representative approaches primarily focus on learning stronger view-invariant representations and alignment mechanisms, such as masked modeling for fine-grained view-invariant video features (BYOV [Park *et al.*, 2025]) or cross-view/cross-modal alignment frameworks (CrossOver [Sarkar *et al.*, 2025]; TransGeo [Zhu *et al.*, 2022]; FG^2 [Xia and Alahi, 2025]). Recent methods further address clutter and drastic appearance shifts: ObjectRelator [Fu *et al.*, 2025] augments segmentation with multimodal conditional fusion and cross-view alignment, while O-MaMa [Mur-Labadia *et al.*, 2025] reformulates correspondence as candidate mask matching to improve zero-shot transferability. These methods mainly rely on segmentation decoders or mask proposal matching, whereas SARA focuses on candidate-level ego–exo retrieval by using region-level semantic anchors as a shared reference frame and fusing anchor-relative topology with language-conditioned visual evidence.

Semantic Promptable Segmentation. In cross-view instance correspondence, masks are the basic carriers for queries and candidates, making semantic-guided segmentation a natural starting point. One line targets general segmentation and directly predicts semantic or instance masks (e.g., Mask R-CNN [He *et al.*, 2017], Mask2Former [Cheng *et al.*, 2022]); another extends to promptable or universal segmentation with point/box/mask or text prompts (e.g., SAM [Kirillov *et al.*, 2023], SAM2 [Ravi *et al.*, 2024], SEEM [Zou *et al.*, 2023], PSALM [Zhang *et al.*, 2025]), often further incorporating language supervision or multimodal models for stronger guidance (e.g., GLIP [Li *et al.*, 2022], LISA [Lai *et al.*, 2024], PixelLM [Ren *et al.*, 2024]). Methods of this type primarily optimize single-view segment quality and generalization, while under cross-view conditions, there may be inconsistency due to viewpoint shifts, because the way prompts and language provide a grounding for regions within a particular image does not establish identity-preserving instance representations across views, as has been indicated by previous work [Tong *et al.*, 2021]. Accordingly, segmentation is best viewed as a high-recall candidate generator for semantic grounding, while cross-view identity reasoning requires additional mechanisms such as semantic anchoring and structured reasoning over the segmented masks.

Instance Mask Matching. Mask matching is the core task of cross-view instance alignment, especially under dense similar instances, occlusions, small targets, and large viewpoint changes [Xie *et al.*, 2024; Lindenberger *et al.*, 2023]. Many studies establish correspondence using scenario-level global representations and overall appearance similarity [Liu *et al.*, 2025], which often fails to achieve fine-grained object-level association [Zhu *et al.*, 2022] under complex local variations, backgrounds, and occlusions. To improve robustness, recent work incorporates geometric cues such as background structure and object relations, leveraging stable scene elements to compensate for viewpoint offsets [Pautrat *et al.*, 2023; Fu *et al.*, 2025]. Other methods achieve this by learning embedding spaces that are independent of viewpoints and modalities or centered around scenes, thereby alleviating the limitations of strict instance-level correspondence and ensuring the consistency of these embedding spaces across different viewpoints and modalities [Sarkar *et al.*, 2025; Xia and Alahi, 2025]. Nevertheless, these methods heavily emphasize global alignment without considering local details or instance changes to morphology. To address this limitation, SARA incorporates local geometric and topological relationships into candidate matching, improving stability and precision under cross-view variations.

3 Methodology

3.1 Task Formulation

This paper considers cross-view instance association in a shared physical scene across two observation spaces: an egocentric view $\mathcal{V}_{ego}$ and an exocentric view $\mathcal{V}_{exo}$. Given an egocentric image $I_{ego} \in \mathbb{R}^{H\times W\times 3}$ and an exocentric image $I_{exo} \in \mathbb{R}^{H'\times W'\times 3}$, the egocentric view is used for user interaction and target specification, while the exocentric view is captured by an external camera or a collaborative system. The target instance O_{tar} in I_{ego} is specified by an open-vocabulary language instruction, based on which a pixel-level target mask $M_{ego} \in \{0,1\}^{H\times W}$ is obtained to delineate the target region. In the exocentric image I_{exo}, a candidate generator (detection/segmentation) produces a candidate set $\mathcal{O} = \{O_1, \ldots, O_N\}$, where each O_i denotes a candidate instance in the exocentric view (implemented as a mask or an index).

This task is designed to simulate a real-world scenario of human-machine collaboration, in which the viewing angle difference between $\mathcal{V}_{ego}$ and $\mathcal{V}_{exo}$ is not subject to any restrictions. We propose a cross-view instance correspondence framework (Figure 1), starting with target specification in the ego view and followed by region extraction in the exo view. The candidate selection process takes into account ROI guidance to maximize recall and reduce computational expense.

3.2 ROI-guided Candidate Mining

To avoid exhaustive matching over the entire exocentric image, a target-guided relevance heatmap $\mathbf{H}$ is first computed to extract a compact region of interest (ROI) $\mathcal{R}_{roi}$. Given the egocentric image I_{ego} and the target mask M_{ego} obtained conditioned on the language instruction, a pretrained visual backbone $\Phi(\cdot)$ produces a dense feature map $\mathbf{F}_{ego}$. Mask-aware pooling over the target region yields a target semantic prototype:

$$\mathbf{v}_{tar} = \mathrm{Pool}(\mathbf{F}_{ego}, M_{ego}). \tag{1}$$

For the exocentric image I_{exo}, the corresponding dense feature map $\mathbf{F}_{exo}$ is extracted, and the heatmap response at each location $p = (x, y)$ is computed by feature similarity:

$$\mathbf{H}(p) = S\big(\mathbf{F}_{exo}(p), \mathbf{v}_{tar}\big), \quad p \in [1, H'] \times [1, W']. \tag{2}$$

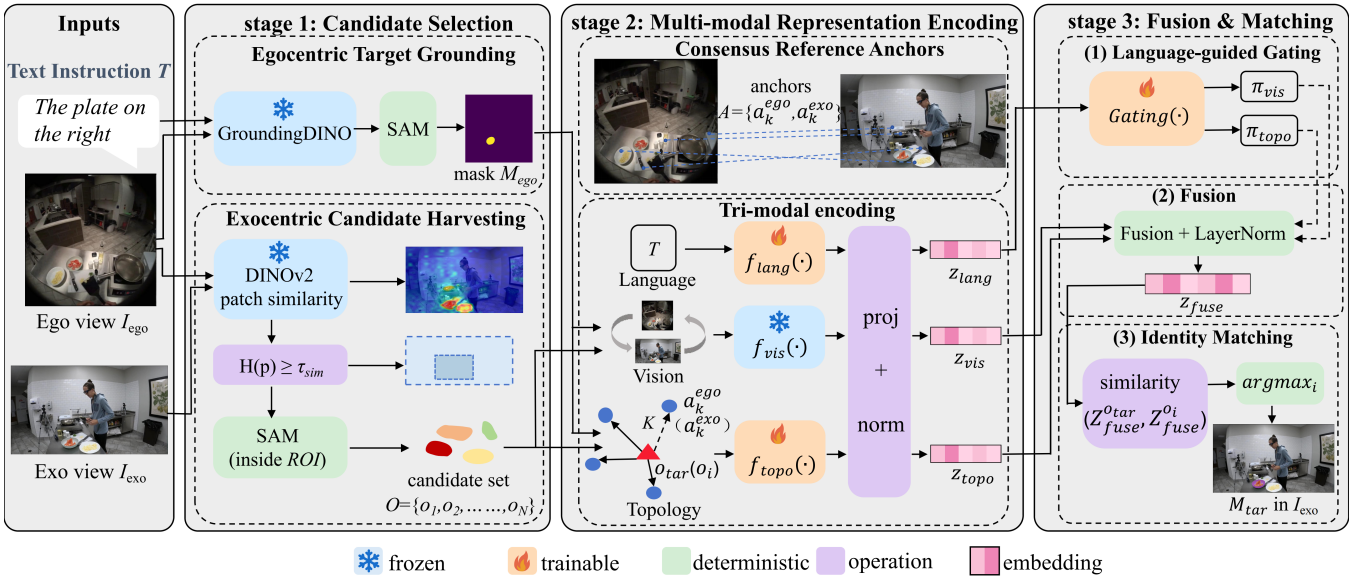

Figure 1: **Overview of SARA**. ROI-guided candidate selection, anchor-based topology encoding, and language-conditioned multimodal fusion for ego–exo instance correspondence.

The ROI $\mathcal{R}_{\text{roi}}$ is obtained by normalizing and thresholding $\mathbf{H}$, followed by a small dilation to cover boundary uncertainty. The segmentation model is applied only within $\mathcal{R}_{\text{roi}}$, rather than over the full exocentric image, to generate candidate masks and form the exocentric candidate set $\mathcal{O} = \{O_i\}_{i=1}^{N}$ for subsequent matching.

3.3 Consensus Reference Anchor Discovery

To reduce identity ambiguity after ROI filtering, we discover a small set of region-level cross-view semantic anchors $\mathcal{A} = \{a_k\}_{k=1}^{K}$ to define a stable referential frame. Each anchor denotes a semantically coherent local region shared across views, rather than an isolated pixel or target candidate. DINOv2 features are used to represent candidate anchor regions, while their centroids are retained only for subsequent topology encoding. To match anchors between views, we compute cosine similarity between the pooled feature representations of ego and exo candidate regions:

$$\mathcal{S}_{\text{anchor}}(A_i^{ego}, A_j^{exo}) = \frac{\langle \mathbf{z}_i^{ego}, \mathbf{z}_j^{exo} \rangle}{\|\mathbf{z}_i^{ego}\|_2 \, \|\mathbf{z}_j^{exo}\|_2}. \tag{3}$$

Mutual-consistency pruning is then applied to retain reliable one-to-one anchor pairs, yielding the final anchor matching set $\mathcal{M}_A = \{(i_k, j_k)\}_{k=1}^{K}$. The resulting shared region anchors provide a consistent cross-view reference frame, whose centroids are used as anchor nodes in the topology graph. When fewer stable anchors are available, the topology branch aggregates over valid anchors, while the final matching relies more on visual and language-conditioned evidence.

3.4 Structured Topology Encoding

After obtaining $\mathcal{A}$, we encode instance–anchor relative geometry into a viewpoint-robust structural representation to complement semantic similarity. For any instance (the ego-side target O_{tar} or an exo-side candidate $O_i \in \mathcal{O}$), a star-shaped topology graph centered at the instance is constructed to explicitly model its spatial context. The construction consists of two steps.

(1) Nodes. We define the center node u_i and anchor nodes $u_{a_k}{}_{k=1}^{K}$. The node set is

$$\mathcal{U}_i = \{u_i\} \cup \{u_{a_1}, \ldots, u_{a_K}\}, \quad |\mathcal{U}_i| = K + 1. \tag{4}$$

The instance location is given by the mask centroid c_i (with c_{tar} for the target), and the anchor locations are a_k (denoted as a_k^{ego} and a_k^{exo} in the corresponding views).

(2) Edges. To reduce the impact of global translation, scale variation, and local projection distortion, only edges between the center node and each anchor node are used, i.e., $\{(u_i, u_{a_k})\}_{k=1}^{K}$. For each edge, the relative displacement is first computed as $\Delta_{i,k} = c_i - a_k$, and a topology transform maps it to a scale-normalized and direction-aligned edge feature $e_{i,k}$. Concretely, a local scale s (the mean distance from anchors to the instance) and a numerical stabilizer ϵ are introduced, and $\Delta_{i,k}$ is decomposed in an anchor-induced local coordinate system $(b, b_\perp)$, where b is the unit vector of the local principal direction and $b_\perp$ is its orthogonal direction. The edge feature consists of the log-distance and normalized projections onto the two orthogonal directions:

$$e_{i,k}^{dist} = \log\Big(\frac{\|\Delta_{i,k}\|_2}{s + \epsilon}\Big), \tag{5}$$

$$e_{i,k}^{x} = \frac{\langle \Delta_{i,k}, b \rangle}{s + \epsilon}, \tag{6}$$

$$e_{i,k}^{y} = \frac{\langle \Delta_{i,k}, b_\perp \rangle}{s + \epsilon}, \tag{7}$$

and is concatenated as

$$e_{i,k} = \mathcal{T}(\Delta_{i,k}) = [e_{i,k}^{dist}, e_{i,k}^{x}, e_{i,k}^{y}] \in \mathbb{R}^3. \tag{8}$$

The log-distance and direction-aligned projections stabilize the topology features against translation and scale changes, reducing local projection-induced drift. This star graph provides a shared structural reference for the target and all exo candidates under the same anchor set.

After graph construction, a lightweight GNN is applied to encode topology relations, improving discriminability while keeping the computational cost controllable. The topology encoder uses a lightweight two-layer message-passing design with mean aggregation to produce the final topology representation of the center node. Let $h_v^{(\ell)}$ denote the node embedding at layer ℓ. Edge-conditioned message passing is performed for the center node u_i: each anchor-node embedding is concatenated with the corresponding edge feature and mapped to a message; messages are aggregated by an order-invariant operator $\mathrm{AGG}(\cdot)$; and the center node is updated by fusing the aggregated message with its current state:

$$m_{i\leftarrow a_k}^{(\ell)} = \phi_\ell([h_{a_k}^{(\ell)} \| e_{i,k}]), \tag{9}$$

$$M_i^{(\ell)} = \mathrm{AGG}(\{m_{i\leftarrow a_k}^{(\ell)}\}_{k=1}^{K}), \tag{10}$$

$$h_i^{(\ell+1)} = \psi_\ell([h_i^{(\ell)} \| M_i^{(\ell)}]). \tag{11}$$

The center-node representation at layer L is taken as the geometric topology vector $\mathbf{g}_i$. Meanwhile, the geometric topology vector explicitly aggregates instance–anchor relative geometry as structural evidence complementary to visual semantics, improving the stability of cross-view representation alignment.

3.5 Language-conditioned Multimodal Manifold Fusion

The candidate set $\mathcal{O} = \{O_i\}_{i=1}^{N}$ is obtained under the ROI constraint, and each candidate instance is associated with a visual semantic vector $\mathbf{v}_i$ and a geometric topology vector $\mathbf{g}_i$. Since these two representations are heterogeneous and typically differ in scale and distribution, naive concatenation or fixed weighting often introduces unstable modality bias under large viewpoint changes and occlusion. Language instructions emphasize key attributes, providing a natural control signal to modulate and fuse visual and topology cues.

The visual semantic representation $\mathbf{v}i$ and the topological representation $\mathbf{g}i$ are both projected into a shared latent space $\mathbb{R}^{d_m}$, so that they can be compared and respectively obtain $\mathbf{z}_{vis,i}$ and $\mathbf{z}_{topo,i}$. The language embedding $\mathbf{z}_{lang}$ is used to compute modality-specific affine modulation parameters: scale γ_m and bias β_m, which are used to reshape the visual and topology features depending on the instruction. These parameters are computed through MLP layers and applied to the features as:

$$\hat{\mathbf{z}}_{m,i} = \gamma_m \odot \mathbf{z}_{m,i} + \beta_m, \tag{12}$$

where $\odot$ denotes element-wise scaling and shifting. This design explicitly exposes instruction-dependent dimension selection: when the instruction stresses appearance attributes (e.g., color or material), modulation increases the contribution of the visual stream; when it stresses spatial relations (e.g., left/right or front/back), modulation strengthens the usability of the topology stream. Computationally, (γ, β) are computed once per instruction, followed by element-wise transformations for each candidate.

After instruction-conditioned modulation, a remaining challenge is to decide which modality is more reliable under the current instruction and scene. To avoid hand-tuned fixed weights (e.g., global hyperparameters $\omega_{sem}, \omega_{topo}$) that may fail across scenarios, a competitive modality routing mechanism is adopted, where the normalized fusion weights are predicted directly from $\mathbf{z}_{lang}$:

$$[\pi_{vis}, \pi_{topo}] = \mathrm{Softmax}(\mathbf{W}_g \mathbf{z}_{lang}), \pi_{vis} + \pi_{topo} = 1. \tag{13}$$

Here, W_g is a learned gating projection that predicts instruction-dependent modality weights. The final multimodal representation for candidate O_i is obtained by a weighted sum of the modulated features, followed by LayerNorm to suppress scale dominance across modalities:

$$\mathbf{z}_i^{mm} = \mathrm{LayerNorm}(\pi_{vis}\hat{\mathbf{z}}_{vis,i} + \pi_{topo}\hat{\mathbf{z}}_{topo,i}). \tag{14}$$

Because fusion entails only combination and normalization for vector-level combinations, the overall computational complexity of fusion has a linear cost of N, allowing for the possibility of performing fusion on many candidate solutions simultaneously. More importantly, $\mathbf{z}_i^{mm}$ unifies appearance and structural evidence in a shared discriminative space, providing a direct input to the probabilistic matching with rejection in the next section.

3.6 Probabilistic Matching with Rejection

Given $\mathbf{z}_{tar}^{mm}$ and $\mathbf{z}_i^{mm}$, matching is formulated as probabilistic retrieval over candidate set. In real-world scenarios, occlusion or limited view may result in the target being absent in I_{exo}. An *empty match* hypothesis $\varnothing$ is introduced to account for this.

The temperature-calibrated cosine similarity is used to compute the matching score between the target and each candidate:

$$s_i = S(\mathbf{z}_{tar}^{mm}, \mathbf{z}_i^{mm}; \tau) \triangleq \frac{1}{\tau} \cdot \frac{\langle \mathbf{z}_{tar}^{mm}, \mathbf{z}_i^{mm} \rangle}{\|\mathbf{z}_{tar}^{mm}\|_2 \, \|\mathbf{z}_i^{mm}\|_2}, \tag{15}$$

$S(\cdot;\tau)$ represents the temperature-calibrated similarity, where the temperature τ controls the score distribution sharpness, helping mitigate over-confidence and ranking instability under viewpoint shifts.

The empty-match hypothesis is added as a constant term in the softmax normalization:

$$\mathcal{P}(O_i \mid O_{tar}) = \frac{\exp(s_i)}{\sum_{k=1}^{N} \exp(s_k) + \exp(\eta)}, \tag{16}$$

where $P(\varnothing|O_{tar})$ is defined analogously by replacing the numerator with $\exp(\eta)$. We predict $\hat{O} = \arg\max_i P(O_i|O_{tar})$ and reject if $P(\varnothing|O_{tar}) > \delta$. This approach performs a single scan over $\mathcal{O}$, unifying matchable and unmatchable cases, and reduces mismatches under extreme viewpoint changes and occlusion. As a result, cross-view instance alignment becomes more robust.

Method	Ego-Exo4D v2 Val (Type-2)			Hard Subset (Type-2)		
	Top-1↑	Recall@5↑	Loc.E↓	Top-1↑	Recall@5↑	Loc.E↓
XSegTx	28.7	42.3	0.084	24.3	38.2	0.089
XMem-XSegTx	44.5	64.2	0.048	39.6	63.3	0.054
DINOv2+kNN (mask pool)	51.2	70.3	0.044	48.8	68.1	0.051
PSALM (zero-shot)	14.3	33.6	0.272	10.9	30.9	0.294
Ours w/o Topology	53.4	76.5	0.042	50.3	71.4	0.045
Ours (SARA)	**71.6**	**92.3**	**0.036**	**69.8**	**91.1**	**0.039**

Table 1: **Main results on Ego-Exo4D v2 val.** We report Top-1, Recall@5, and Loc.E on the Type-2 full set and its hard subset; all methods use the same oracle exo-view candidates (GT proposals), isolating cross-view matching/retrieval accuracy from candidate-generation quality.

4 Experiments

4.1 Experimental Setup

Dataset and Split. The proposed method is evaluated on the **Ego-Exo4D Correspondences (v2)** benchmark, which provides time-synchronized first-person (FPV, ego) and third-person (TPV, exo) video observations, along with instance-level annotations for cross-view instance association and collaborative perception research.

To ensure reproducibility and verifiability, we focus on task-relevant scene stratification on v2 val, targeting the subset of matchable collaborative scenes. The model is trained using the official v2 training set (train), while the scene stratification and primary evaluation are conducted solely on v2 val to ensure fairness. The evaluation unit is defined as the time-synchronized ego–exo frame pair. Since the goal is ego–exo instance correspondence rather than mask generation, the main evaluation uses oracle exo-view GT instance masks as an identical candidate set for all methods, thereby measuring matching/retrieval accuracy independently of proposal quality.

Scenario Stratification. The Ego-Exo4D dataset spans a wide range of scenes, not all of which align with the task assumption of "language-referable target + segmentable instance + cross-view collaborative localization". We stratify the scenes in **v2 val (total 842 scenes)** based on object referability, cross-view visibility, and candidate interference. The following three categories are defined:

1. **Type-1 (Trivial single-instance)**: In these scenes, both the ego and exo views contain only a single prominent candidate, with negligible ambiguity in matching (248 scenes). 2. **Type-2 (Matchable collaborative scenes)**: These scenes have corresponding target instances in both ego and exo views, with multiple candidate interference in the exo view. This represents the core cross-view matching scenario of this work (214 scenes). 3. **Type-3 (Non-collaborative / non-referable)**: Scenes where no stable language-referable or segmentable objects are present (e.g., pure human motion, rock climbing), which fall outside the scope of this collaborative localization task (380 scenes).

The dataset is divided into three categories: Type-1 for correctness checks, Type-2 for evaluation, and Type-3 for failure case analysis.

Candidate-based Protocol and Candidate Pool. While Ego-Exo4D provides instance-level cross-view annotations, this work adopts a **candidate-based identity association** protocol. Given a pair of time-synchronized ego-exo frames, the target instance is specified by the ego-view mask M_{ego} (with an optional language instruction T), and the model must select the corresponding target instance from a candidate set $\mathcal{O} = \{O_1, \ldots, O_N\}$ in the exo view.

To ensure fair comparison, an **oracle candidate pool** (GT proposals) is adopted for the exo-view candidate set. The number of candidates N varies with the number of annotated instances in the exo view, i.e., a variable-length candidate set. Evaluation metrics are calculated across these variable-length candidate sets, with Recall@5 ensuring comparability for smaller sets. Segmentation-based methods are ranked based on IoU with each candidate mask to ensure fair comparison.

To further evaluate robustness under challenging conditions, we construct a **hard subset** within the **Type-2** main evaluation set, where sample selection is based solely on dataset annotations and does not depend on model predictions. The challenging conditions include similar instance density, small target size, and occlusion.

Metrics and Implementation. The performance of cross-view instance matching is evaluated using three core metrics: **Top-1 Accuracy**, **Recall@5**, and **Localization Error (Loc.E)**. Top-1 and Recall@5 measure identity matching and candidate ranking quality, while Loc.E is the normalized Euclidean distance between the predicted and ground truth mask centroids. For $N < 5$, Recall@5 is computed over the top-$\min(5, N)$ candidates. DINOv2 is used to extract dense features and instance-level descriptors, while the geometric branch discovers shared anchors and encodes spatial relations with a lightweight GNN. Multimodal fusion is performed through language-conditioned modulation.

4.2 Main Results

Baseline Models. The compared baselines fall into three representative paradigms:

(i) official cross-view segmentation models with spatiotemporal memory (e.g., XSegTx and XMem-XSegTx), which propagate query-view information to the other view via mask prediction or memory transfer;

(ii) a purely retrieval-based visual baseline (DINOv2 + kNN), which ranks exo candidates by appearance similarity;

(iii) a language-driven zero-shot segmentation paradigm (PSALM), which emphasizes open-vocabulary generaliza-

Method	Top-1↑	Recall@5↑	Loc.E↓
GroundingDINO + SAM	43.8	69.1	0.082
PSALM (zero-shot)	9.2	18.5	0.401
Ours (SARA)	**63.4**	**87.2**	**0.041**

Table 2: **Text-conditioned candidate-based evaluation on Ego-Exo4D v2 Val (Type-2 Full).** Each method localizes the ego target from text and retrieves the exo match from the same oracle exo-view candidates (GT proposals), evaluating language-conditioned matching/retrieval rather than proposal quality. Metrics: Top-1, Recall@5, Loc.E.

tion from text to masks. This selection follows common practice in Ego–Exo correspondence evaluation, covering both strong retrieval baselines and segmentation/LLM-seg paradigms, and enables disentangling improvements from cross-view identity ranking versus mask generation capability.

Comparison on Type-2 Full. As shown in Table 1, SARA significantly outperforms all baselines across all metrics on the Type-2 Full set. Notably, SARA improves Top-1 accuracy by 20.4 points over the strongest visual-retrieval baseline (71.6% vs. 51.2%). This performance gap is further elucidated by the comparison with our "Ours w/o Topology" variant: compared with the ROI-guided visual variant, full SARA improves Top-1 by 18.2 points (71.6% vs. 53.4%). These results quantitatively demonstrate that under extreme viewpoint shifts and candidate interference, local appearance cues—even those from powerful pre-trained models like DINOv2—are insufficient for reliable identity association. By integrating language-conditioned retrieval with structured geometric priors, SARA effectively resolves such ambiguities, leading to both superior ranking quality and more precise localization (Loc.E of 0.036).

Comparison on the Hard Subset. On the Hard Subset, all methods degrade to varying degrees, while SARA exhibits a smaller drop and a stronger relative margin in Table 1. This supports that the hard cases are dominated by identity confusion induced by dense distractors, high instance similarity, and occlusions—exactly the bottlenecks targeted by SARA. In these scenes, retrieval by appearance is more likely to be misled by near-duplicates, and segmentation/memory-based baselines are also prone to drifting under cross-view distortions and partial visibility. In contrast, SARA benefits from topology encoding that provides stable relational constraints among similar instances, ROI guidance that reduces ranking noise under cluttered candidate sets, and language-conditioned modulation that strengthens referential cues when visual evidence is weak. As a result, SARA maintains higher Top-1/Recall@5 and keeps Loc.E low under challenging collaborative conditions.

Text-conditioned Candidate-based Evaluation. Table 2 evaluates a text-conditioned setting, where the ego target is specified by language and the exo match is retrieved from the same oracle candidate set. The text-conditioned setting is not directly comparable with the mask-specified protocol, as it additionally includes ego-target grounding noise. This setting extends the main mask-specified protocol by testing whether language cues can guide target grounding and cross-view retrieval under the same candidate-based evaluation. SARA outperforms GroundingDINO + SAM and PSALM, indicating that language-conditioned fusion and anchor-based topology provide complementary evidence beyond text-driven grounding alone. These results further show that SARA remains effective when the target specification comes from natural language rather than a predefined ego mask.

4.3 Ablation Studies

Goal and Variants. The progressive ablation (A0–A4) is conducted on Ego-Exo4D v2 Val (Type-2 Full) to isolate the effect of each component under the same evaluation setting. All variants share the *oracle exo-view candidate pool* (GT proposals) and the unified candidate-based retrieval protocol. A0 ranks candidates using visual-only instance representations. A1 adds ROI pruning that *does not* alter the oracle pool, but only masks/penalizes scores for out-of-ROI candidates. A2 incorporates topology cues with non-learned handcrafted aggregation, while A3 replaces it with a learnable GNN encoder for relational geometry. A4 is the full model with language-conditioned multimodal modulation and fusion on top of A3.

Results and Analysis. Clear additive gains are observed as constraints and structured cues are introduced in Table 3. ROI pruning delivers the largest single-step improvement (A0→A1), indicating that irrelevant candidates dominate ranking noise in Type-2 scenes. Topology further improves both retrieval and localization (A1→A2→A3), confirming that relational geometry provides complementary evidence beyond appearance, especially under viewpoint distortion and partial visibility. Replacing handcrafted pooling with a learnable GNN yields consistent gains, suggesting the benefit of adaptive relational modeling. Finally, language-conditioned fusion brings a substantial boost (A3→A4), indicating that language helps to optimize the allocation of weights for visual and geometric features by emphasizing the relevant attributes of the target. All the results validate the necessity of ROI denoising, topology-based disambiguation, and language modulation for cross-view identity association.

4.4 Robustness in Collaboration Scenarios

To assess cross-scenario generalization under realistic collaboration settings, robustness testing is conducted on a self-collected ego–exo collaboration dataset. The dataset includes synchronized ego–exo video pairs, language instructions, and instance annotations for evaluation of target localization. We use this set only as an auxiliary robustness check rather than the primary benchmark. Samples are grouped into three subsets—multi-object, high-similarity, and occlusion scenes—and evaluated with Top-1, Recall@5, and Loc.E.

As shown in Table 4, SARA consistently achieves the best performance across all three scenarios, with higher Top-1/Recall@5 and lower Loc.E than the baselines, indicating more accurate cross-view association and more stable lo-

Ablation Variant (Progressive)	Top-1↑	Recall@5↑	Loc.E↓	**Gain**
A0: Visual-only	40.3	53.8	0.083	–
A1: A0 + ROI pruning	53.4	76.5	0.042	13.1
A2: A1 + Topology (non-learned pooling)	57.1	78.4	0.040	3.7
A3: A2 + Topology (GNN)	62.3	80.2	0.039	5.2
A4: Full SARA (A3 + Language-conditioned fusion)	**71.6**	**92.3**	**0.036**	**9.3**

Table 3: **Progressive ablation study on Ego-Exo4D v2 Val (Type-2 Full).** A0–A4 progressively add ROI pruning, topology encoding, and language-conditioned fusion. **Gain**: Top-1 improvement over the previous variant.

Method	Multi-object Scenes			High Similarity Objects			Occlusion Scenes		
	Top-1↑	Recall@5↑	Loc.E↓	Top-1↑	Recall@5↑	Loc.E↓	Top-1↑	Recall@5↑	Loc.E↓
GroundingDINO + SAM	37.5	72.4	0.091	34.2	64.3	0.103	39.8	74.9	0.084
PSALM (zero-shot)	9.1	17.8	0.418	3.5	10.9	0.582	5.6	16.8	0.497
Ours (SARA)	**61.4**	**85.3**	**0.048**	**55.1**	**78.9**	**0.063**	**59.7**	**84.4**	**0.060**

Table 4: **Robustness testing on self-collected dataset.** Results are reported on three challenging subsets (multi-object, high-similarity, and occlusion scenes).

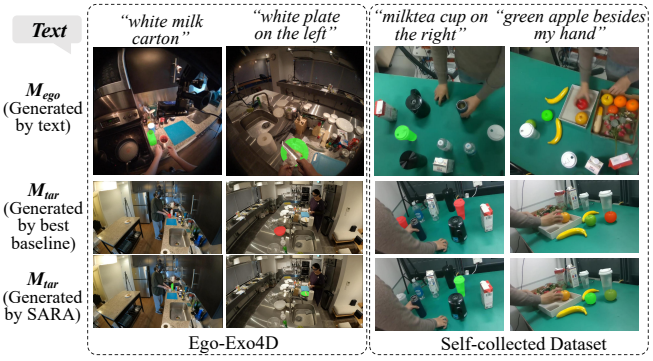

Figure 2: **Qualitative results on Ego-Exo4D and self-collected dataset.**

calization. In multi-object scenes, SARA maintains reliable ranking, showing resilience to candidate interference. In high-similarity scenes, SARA outperforms zero-shot segmentation baselines, benefiting from topology constraints and language modulation to disambiguate similar instances. In occlusion scenes, performance degradation remains relatively controlled, indicating that multimodal evidence fusion provides more stable decision making under partial visibility.

4.5 Qualitative Results

To better understand common failure modes in cross-view instance association and the mechanisms behind SARA, qualitative visualizations are provided on Ego-Exo4D Type-2 scenes shown and self-collected dataset in Figure 2. Two dominant baseline errors are observed: *(i) similar-instance attraction* in dense scenes, and *(ii) insufficient local evidence* under occlusion and viewpoint distortion. Thus, visual retrieval may favor nearest neighbors, while zero-shot segmentation may drift to related regions.

In contrast, SARA exhibits a more stable decision pathway across views. ROI pruning removes irrelevant candidates and stabilizes ranking under dense interference. Topology adds relational constraints beyond appearance, reducing look-alike confusion. Language-conditioned fusion emphasizes referential attributes, improving instruction alignment. The visualizations show that SARA consistently localizes the correct instance even under small targets, strong occlusion, and complex backgrounds. These observations support the complementarity of ROI, topology, and language-conditioned fusion.

5 Conclusion

In this paper, we have addressed the critical challenge of ego–exo instance correspondence, a fundamental bottleneck for seamless human-machine collaboration. To mitigate the severe semantic drift and identity confusion caused by drastic viewpoint shifts and occlusions, we proposed SARA (Semantic-Anchored Referential Alignment). By leveraging viewpoint-invariant semantic anchors, our framework establishes stable reference points and provides interpretable structural evidence through ROI-guided denoising and topological modeling. Furthermore, the integration of language-conditioned modulation and multimodal fusion enables SARA to achieve adaptive alignment with user intent. Extensive evaluations on Ego-Exo4D, its hard subsets, and self-collected datasets consistently demonstrate SARA's superiority in matching precision, ranking robustness, and localization stability. Our results confirm that the complementary fusion of visual, topological, and linguistic modalities is a key pathway toward achieving robust cross-view synergetic localization in unconstrained environments.

Acknowledgments

This work was supported in part by the NSF of China under Grant 62473251, and in part by the National Key Research and Development Program of China under Grant 2024YFB3312100.

References

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.

[Fu *et al.*, 2025] Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6530–6540, October 2025.

[Grauman *et al.*, 2024] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, June 2024.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[Lai *et al.*, 2024] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, June 2024.

[Li *et al.*, 2022] 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 (CVPR)*, pages 10965–10975, June 2022.

[Lindenberger *et al.*, 2023] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed, 2023.

[Liu *et al.*, 2025] Tao Liu, Kan Ren, and Qian Chen. Object detection as an optional basis: A graph matching network for cross-view uav localization, 2025.

[Mur-Labadia *et al.*, 2025] Lorenzo Mur-Labadia, Maria Santos-Villafranca, Jesus Bermudez-Cameo, Alejandro Perez-Yus, Ruben Martinez-Cantin, and Jose J. Guerrero. O-mama: Learning object mask matching between egocentric and exocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6892–6903, October 2025.

[Park *et al.*, 2025] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bootstrap your own views: Masked ego-exo modeling for fine-grained view-invariant video representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13661–13670, June 2025.

[Pautrat *et al.*, 2023] Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. Gluestick: Robust image matching by sticking points and lines together. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9706–9716, October 2023.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[Ren *et al.*, 2024] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26374–26383, June 2024.

[Robinson *et al.*, 2023] Nicole Robinson, Brendan Tidd, Dylan Campbell, Dana Kulić, and Peter Corke. Robotic vi-

sion for human-robot interaction and collaboration: A survey and systematic review. *J. Hum.-Robot Interact.*, 12(1), February 2023.

[Sarkar *et al.*, 2025] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8985–8994, June 2025.

[Sun *et al.*, 2025] Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. *arXiv preprint arXiv:2504.13621*, 2025.

[Tong *et al.*, 2021] Xin Tong, Xianghua Ying, Yongjie Shi, He Zhao, and Ruibin Wang. Towards cross-view consistency in semantic segmentation while varying view direction. In *IJCAI*, pages 1054–1060, 2021.

[Xia and Alahi, 2025] Zimin Xia and Alexandre Alahi. Fgˆ2: Fine-grained cross-view localization by fine-grained feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6362–6372, June 2025.

[Xie *et al.*, 2024] Yaxu Xie, Alain Pagani, and Didier Stricker. Sg-pgm: Partial graph matching network with semantic geometric fusion for 3d scene graph alignment and its downstream tasks, 2024.

[Zhang *et al.*, 2022] Baoquan Zhang, Shanshan Feng, Xutao Li, Yunming Ye, Rui Ye, Chen Luo, and Hao Jiang. Sgmnet: Scene graph matching network for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[Zhang *et al.*, 2025] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 74–91, Cham, 2025. Springer Nature Switzerland.

[Zhu *et al.*, 2022] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1162–1171, June 2022.

[Zou *et al.*, 2023] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19769–19782. Curran Associates, Inc., 2023.