ARL-TR-10183 ● SEP 2025

Multimodal 3D Gaze–Object Localization for Shared Human–Machine Situational Awareness in Tactical Environments

by Soorya Saravanan and Russell Cohen Hoffing

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10183 ● SEP 2025

Multimodal 3D Gaze–Object Localization for Shared Human–Machine Situational Awareness in Tactical Environments

Soorya Saravanan
Army Educational Outreach Program HBCU-MI Summer Research Program, University of California, Riverside

Russell Cohen Hoffing
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
September 2025	Technical Report	06/02/2025	08/07/2025

4. TITLE AND SUBTITLE
Multimodal 3D Gaze–Object Localization for Shared Human–Machine Situational Awareness in Tactical Environments

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Soorya Saravanan and Russell Cohen Hoffing

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FC Adelphi, MD 20783	ARL-TR-10183

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Modern military operations require seamless human–machine integration to enhance situational awareness. This research, part of the Distributed Information for Enhanced Squad Lethality (DIESL) program, introduces a novel 3D gaze–object localization system that creates machine-readable human-generated environmental information for shared squad awareness without manual input from users. The system fuses multimodal data, from including Soldier eye-tracking, first-person 2D video feeds, camera pose, and visual embeddings from the CLIP vision-language model, to uniquely identify objects observed by different team members across time and space. The central hypothesis is that combining geometric data with visual features resolves ambiguities inherent in computer vision-only approaches. To test this, the system was evaluated using data from field experiments with Soldiers. Results demonstrate that fusing geometric and visual features significantly improves gaze–object localization accuracy compared to geometric-only baselines. The resulting 3D localized objects enable human operators and autonomous systems to share a synchronized understanding of the environment, thus improving coordination and mission effectiveness. This work validates multimodal sensor fusion from opportunistically sensed human-centric data as foundational for future advances in human–machine teaming.

15. SUBJECT TERMS
Humans in Complex Systems, human–machine integration, situational awareness, multimodal sensor fusion, opportunistic sensing, eye tracking, computer vision

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	22

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Russell Cohen Hoffing	(520) 691-5011

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Multimodal sensing platforms are increasingly used to support human–machine teaming in tactical environments by capturing gaze, pose, and environmental features from both Soldiers and autonomous systems. In these settings, head-mounted sensors and body-worn systems can provide real-time insight into what individual operators are observing; thus, enabling improved coordination and shared situational awareness. While previous work has explored geometric gaze estimation (Mora and Odobez 2014) and object detection (Zhao et al. 2019) separately, few efforts have combined these modalities to achieve persistent, object-level gaze localization in three dimensions (Kellnhofer et al. 2019). In particular, current methods often struggle with ambiguity when multiple Soldiers observe similar objects from different viewpoints (Beyerer et al. 2021), or when sensor noise degrades geometric accuracy (Zemblys et al. 2019).

In this research, we propose a method for localizing the intersection of a Soldier's gaze with objects in a 3D environment using both geometric and visual features. Our system fuses gaze vectors, camera pose data, and high-dimensional visual embeddings to identify and persistently associate gaze hits with objects over time. This fused representation allows multiple users—and autonomous platforms such as unmanned aerial systems (UAS)—to share a consistent, machine-readable map of which targets have been observed, when, and by whom.

2. Methods

Data was collected during controlled-field and mixed-reality experiments using a suite of wearable and environmental sensors. Each participant was equipped with a Microsoft HoloLens 2 headset, which provided first-person video, eye tracking, and head-pose tracking to capture gaze and head position and orientation. Participants walked freely throughout the environment and were tasked with visually identifying and aiming their gaze at static-colored targets—specifically, red and yellow boxes positioned throughout the space. The 3D positions of these targets were recorded by standing at each box location with the HoloLens 2 to capture accurate spatial ground truth. All data streams were time stamped and synchronized to ensure accurate temporal alignment for downstream analysis.

2.1 Data Standardization with STAIT

All raw data collected from the HoloLens including gaze, head pose, and video were converted into the Spatiotemporal Analysis and Inference Toolkit (STAIT) format. STAIT addresses the lack of standardization in coordinate systems across

VR/AR/XR devices and sensors. For example, axis definitions such as the positive Y-axis may vary across platforms, which complicates multisensor fusion and spatial alignment.

STAIT provides a unified coordinate system and structured data schema for representing 3D position, orientation, and movement over time. Inspired by standards like the Brain Imaging Data Structure and International Society of Biomechanics recommendations, STAIT is designed for use in multiagent mixed-reality environments. It enables seamless integration of sensor data from multiple humans and autonomous systems operating concurrently. Standardizing all HoloLens data into STAIT ensures that downstream processing, visualization, and model development are robust, reproducible, and generalizable across platforms.

2.2 Data Preprocessing and Cleaning

Once standardized into the STAIT format, raw sensor streams—including gaze vectors, camera pose, and object detections—were passed through a multistage preprocessing pipeline. Object detections were originally generated from first-person video frames using an in-house–developed Grounding DINO object detector (Parnerkar 2024), which produced bounding boxes around visible targets using the prompts "red box" and "yellow box." Only detections with confidence scores above a defined threshold (typically 0.5) were retained, ensuring that the pipeline focused on high-quality visual signals.

Denoising was performed by filtering out unreliable or misaligned samples. Specifically, gaze points were kept only if they fell within one of the high-confidence bounding boxes in the corresponding frame. This step eliminated noisy gaze–object pairings and ensured that only plausible interactions were retained. Importantly, this gaze–object pairing ensured ground truth matching where the 3D gaze that was paired with the gaze–object (extracted from the 2D image) would be used to estimate the 3D location of the object. Additional filtering excluded frames with insufficient data, such as those occurring before a defined starting index or lacking synchronized sensor metadata.

Following this filtering process, the remaining data were temporally aligned so that each frame included valid and synchronized gaze, camera pose, and object detection information. Only frames where the gaze vector intersected a detected object were used in downstream training and evaluation.

Visual inspection tools were applied to confirm accurate alignment between gaze and detection data. For example, Figure 1 shows a representative case from Frame 1063, where a red bounding box identifies a detected object and a blue dot marks

the corresponding 2D gaze point. This overlay confirms proper spatial and temporal synchronization across modalities.

Figure 1. Gaze and object detection alignment red box: DINO-detected object; blue dot: projected 2D gaze. Shows successful gaze–object pairing.

Frames showing spatial inconsistencies—such as gaze points clearly missing all nearby high-confidence detections, or cases of detection misalignment—were excluded from the final dataset.

2.3 Gaze Intersection Logic and Feature Engineering

A core component of this pipeline is the gaze intersection logic, which maps 2D gaze vectors to 3D world coordinates. This is done by modeling gaze as a 3D ray originating from the eye and extending outward along the gaze direction. The ray is parameterized as

$$P(t) = P_0 + t \cdot d \quad \text{P(t) = Po+t*d} \tag{1}$$

where P_0 is the 3D gaze origin and d is the normalized gaze direction vector. To estimate the 3D point of regard, this gaze ray is intersected with a virtual ground plane defined dynamically for each frame. The ground plane is offset from the HoloLens frustum origin using an average person's height, and a learned correction vector ($\Delta x = -0.243$, $\Delta y = -0.821$, $\Delta z = 0.642$) is applied to compensate for consistent sensor bias observed during empirical calibration. This vector was derived through statistical analysis of gaze intersection errors and manual tuning based on ground truth object positions—correcting for forward, lateral, and vertical

offsets caused by head-pose estimation inaccuracies, eye-tracking asymmetries, and floor-level assumptions. Object bounding boxes, originally detected in 2D image space by the Grounded DINO model, are filtered based on confidence score and alignment with gaze rays. Only gaze points that fall inside high-confidence bounding boxes are retained for intersection computation. The 3D location of the corresponding object on the ground is then used as the target point for model training and evaluation.

Feature vectors were constructed to represent both geometric and semantic context. Geometric features included the 3D gaze origin and direction, head and camera pose, and the 2D pixel coordinates of the gaze point on the red, green, and blue (RGB) image.

For the Contrastive Language-Image Pretraining (CLIP)-augmented model, visual features were extracted using the CLIP model. Two embeddings were generated per frame: one from the entire RGB scene, and another from a tight crop around each detected object. Each crop was passed through CLIP to obtain a 512-dimensional embedding. These embeddings were concatenated with the geometric features to form a single multimodal input vector.

In contrast, the classic model used only geometric features. Model performance was evaluated using Euclidean distance between the predicted and ground truth object positions, as well as error distributions across frames and object instances to assess accuracy and robustness.

2.4 Model Training and Evaluation

Three machine learning (ML) pipelines were developed and evaluated:

1) Model 1: Raw Geometric (M1-RG) is a nonlearning baseline that estimates 3D gaze–object intersections through direct geometric projection. It computes the intersection point (as described above) by extending a normalized gaze vector from the eye origin and calculating its intersection with the ground plane. This method does not involve any training or optimization and relies solely on raw gaze direction and head-pose data (camera position and orientation). As a result, it provides a reference point for evaluating the limitations of uncorrected sensor inputs, particularly in real-world noisy environments.
2) Model 2: Classic Model (M2-CM) introduces a data-driven approach by training a Random Forest Regressor on a set of 18 geometric features. These features include the 3D origin and direction of the gaze vector, head- and camera-pose vectors, and the pixel coordinates of the gaze point on the RGB

frame. The model is trained using supervised learning and designed to correct for systematic biases, sensor drift, and projection errors that are common in raw geometric estimations. Its architecture is based on an ensemble of decision trees, which partitions the feature space to make robust nonlinear predictions of the 3D intersection point.

3) Model 3: CLIP-Augmented Model (M3-CLIP) extends the M2-CM pipeline by incorporating semantic visual context through CLIP-based image embeddings. For each frame, 512-dimensional embeddings are extracted using the CLIP model, applied to both the full-scene RGB image and a cropped region surrounding the detected object. These two embeddings are concatenated and subsequently reduced via Principal Component Analysis (PCA). Following dimensionality reduction, univariate feature selection is performed using the SelectKBest method with f_regression to isolate the four most predictive dimensions. These four CLIP-derived features are then concatenated with the original 18 geometric features to form a 22-dimensional multimodal input vector. This combined input is used to train the same Random Forest Regressor architecture as in M2-CM, allowing the model to leverage both spatial and semantic cues.

Both M2-CM and M3-CLIP undergo hyperparameter optimization through a grid search procedure, with parameters such as the number of estimators, maximum tree depth, and minimum samples per leaf tuned using five-fold cross-validation. This ensures that the models generalize effectively to unseen data and are robust to overfitting.

The models are subsequently evaluated using consistent performance metrics across all approaches. Each model is assessed using Mean Absolute Error (MAE); average magnitude of prediction error; Root Mean Squared Error (RMSE), which penalizes larger errors to reflect error variance; and Median Error, which is less influenced by outliers.

We compared all three pipelines—M1-RG (baseline), M2-CM (geometric learning), and M3-CLIP (multimodal fusion)—and analyzed error histograms and residual plots to interpret model robustness and pinpoint sources of failure.

3. Results

This section presents the quantitative and qualitative results of the 3D gaze–object localization system across three modeling approaches: M1-RG, M2-CM, and M3-CLIP. For each approach, results are shown for both Participant 1 and Participant 2, with side-by-side visualizations and corresponding error metrics. Although

labeled as Run 1 and Run 2 in the dataset, both refer to the same experiment conducted with two different participants (Table 1).

Table 1. Quantitative metrics of each model per run.

Model	MAE (m)	RMSE (m)	Std. error (m)	Median error (m)	N points
M1-RG Run 1	2.512	2.998	1.644	1.945	104
M1-RG Run 2	2.595	5.292	3.911	2.524	71
M2-CM Run 1	0.350	0.972	0.915	0.000	104
M2-CM Run 2	0.648	1.296	1.229	0.000	71
M3-CLIP Run 1	0.343	0.961	0.898	0.000	104
M3-CLIP Run 2	0.551	1.348	1.230	0.000	71

The raw geometric approach yields the highest localization errors, with MAEs exceeding 2.5 m for Participant 1 and 3.5 m for Participant 2 as seen in Table 1. While these errors may appear large in absolute terms, they are often acceptable for general situational awareness tasks, such as helping a teammate locate an object or region of interest on a shared map. However, visual inspection shows that many predicted gaze–object intersections still fall noticeably away from the ground truth, pointing to the limitations of using uncorrected geometric estimation when more precise localization is needed (Figure 2).

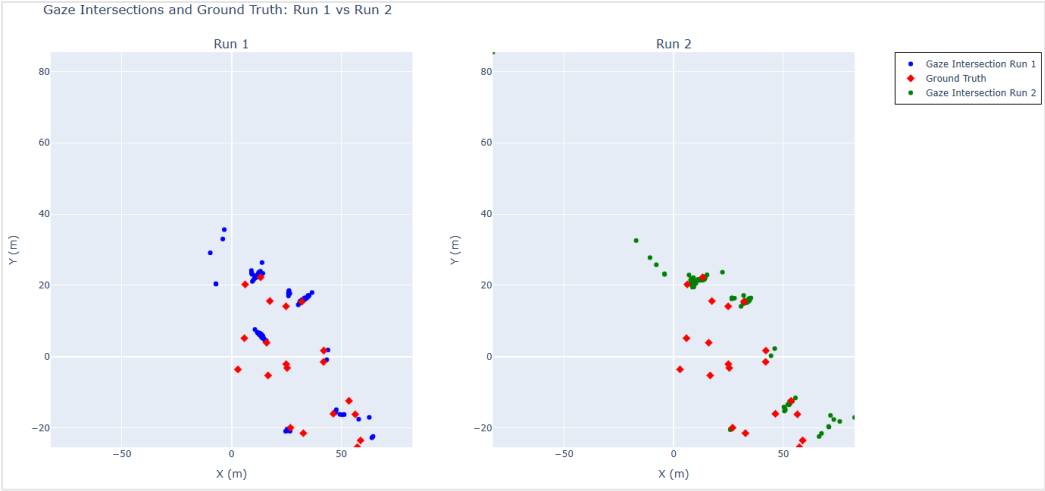

Figure 2. Raw geometric predictions for Participant 1 and Participant 2, plotted alongside ground truth object locations. The gaze intersection points, computed solely from geometric calculations without any learning or correction (M1-RG), exhibit substantial deviation from the ground truth for both participants.

3.1 Classic Model

Model M2-CM significantly improves upon M1-GM, reducing MAE to under 0.5 m in both runs as seen in Table 1. A median error of zero in both cases indicates that over half of the predictions are precisely aligned with ground truth. Visually, the predicted points are tightly clustered around true object positions, demonstrating the effectiveness of geometric feature fusion combined with ML-based correction (Figure 3).

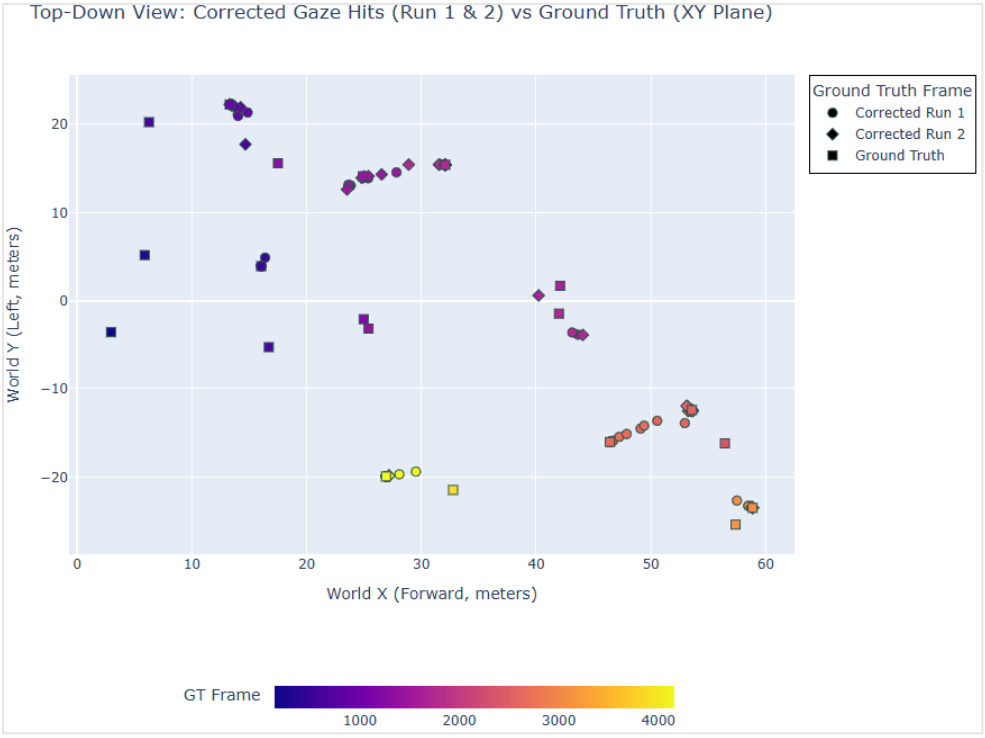

Figure 3. Classic model predictions for both runs, plotted side by side with the ground truth object locations. The model fuses multiple geometric features and applies a Random Forest regressor to learn systematic corrections from data. No visual (CLIP) features are used in this version.

3.2 CLIP-Augmented Model

Model M3-CLIP is a dramatic improvement over the raw geometry model. Relative to M1-GM (MAE 2.512 and 3.595 m), M3-CLIP achieves an 86% reduction in MAE for Run 1 and 85% reduction for Run 2, demonstrating that visual embeddings powerfully compensate for sensor noise and ambiguous gaze rays. In Run 1, M3-CLIP (MAE 0.343 m) slightly outperforms M2-CM (MAE 0.350 m) as seen in Table 1, showing that semantic context can deliver better correction to purely geometric learning. The zero median error in both runs indicates that at least half of all predictions are spot-on; the CLIP components help tighten the tails of the error distribution, as reflected in the respectable RMSE and standard error values.

These results confirm that CLIP-derived features offer substantial and robust benefits when combined with geometric data, especially in noisy or visually complex scenarios. With further tuning of feature fusion weights and architectures, we anticipate CLIP augmentation will consistently surpass the classic model and drive localization accuracy further downward (Figure 4).

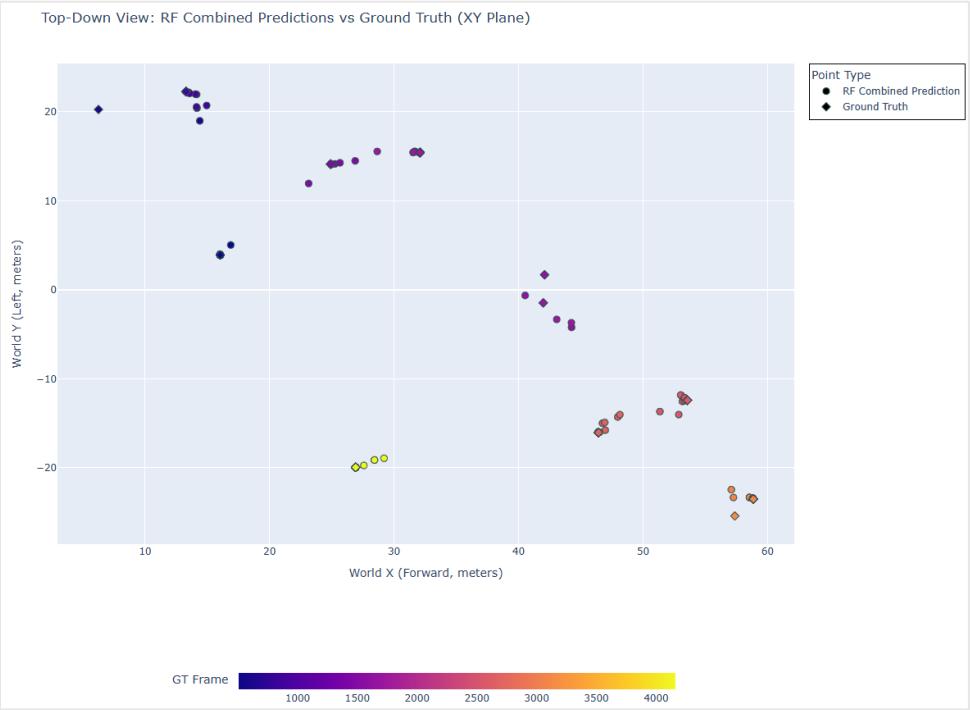

Figure 4. Predictions of the CLIP-augmented model for both runs, again shown side-by-side with ground truth. This model incorporates both geometric features and CLIP-derived visual embeddings to leverage additional scene and object context.

3.3 Clustering in the Raw Gaze–Object Localization Pipeline

The clustering system employs DBSCAN (Density-Based Spatial Clustering of Applications with Noise) as the primary algorithm for grouping gaze–object predictions across three model outputs to do object localization because each gaze–object point should reflect a single location of the predicted static object. DBSCAN was selected over alternatives like K-means or hierarchical clustering for several compelling reasons. First, DBSCAN does not require pre-specifying the number of clusters, which is essential since the number of distinct objects in a scene varies dynamically across frames and runs. Second, DBSCAN can identify clusters of arbitrary shapes and sizes, which is critical in 3D spatial contexts where gaze–ray intersections may not form regular or spherical clusters. Third, DBSCAN is designed to handle noise by assigning outlier points a cluster ID of –1, which is beneficial for discarding spurious or ambiguous gaze–object intersections.

Each pipeline uses DBSCAN in a way that reflects the type of features available: In the raw pipeline, clustering operates on direct geometric intersection points (x, y, z) computed from ray casting gaze vectors into the scene. Although visual features are merged to assist matching, the clustering uses cosine distance on CLIP embeddings to group gaze hits that likely correspond to the same physical object (Figure 5).

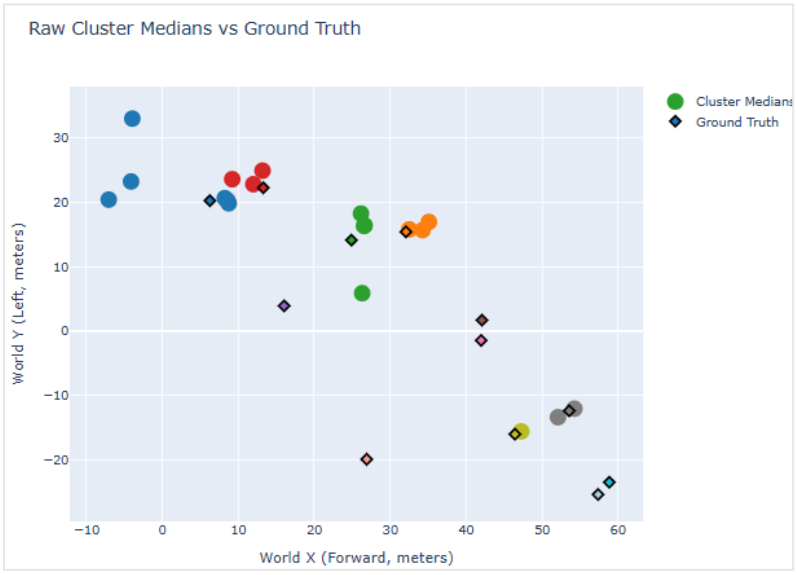

Figure 5. (Raw): Clustering on geometric gaze–ray intersection group predictions based on physical proximity, resulting in basic but direct cluster structures.

In the RF pipeline, clustering is applied to predicted object locations generated by a trained Random Forest model. These predictions incorporate learned relationships between geometric features and true object locations. Here again, CLIP embeddings are used for clustering, and the cosine distance metric ensures that gaze predictions for visually similar objects are grouped together, even if slightly offset in space (Figure 6).

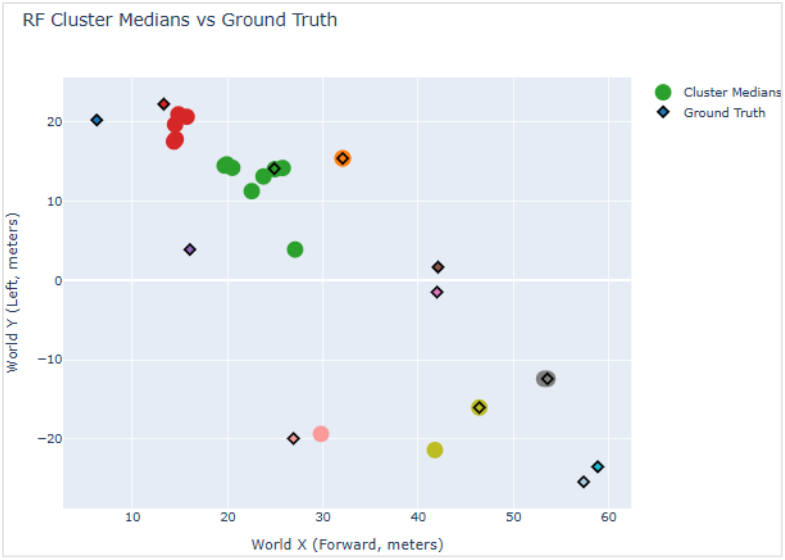

Figure 6. (RF): Clustering on RF predictions shows more spatially coherent groupings due to learned geometric correction.

In the CLIP pipeline, predictions are enhanced with both geometric and visual embeddings. CLIP features extracted from the scene and object crops are processed with dimensionality reduction and feature selection (PCA + SelectKBest) and then standardized before being passed into the model. Clustering is again performed on the output predictions using the associated CLIP embeddings, leveraging semantic similarity across frames and viewpoints to group multiple predictions that are likely to correspond to the same object (Figure 7).

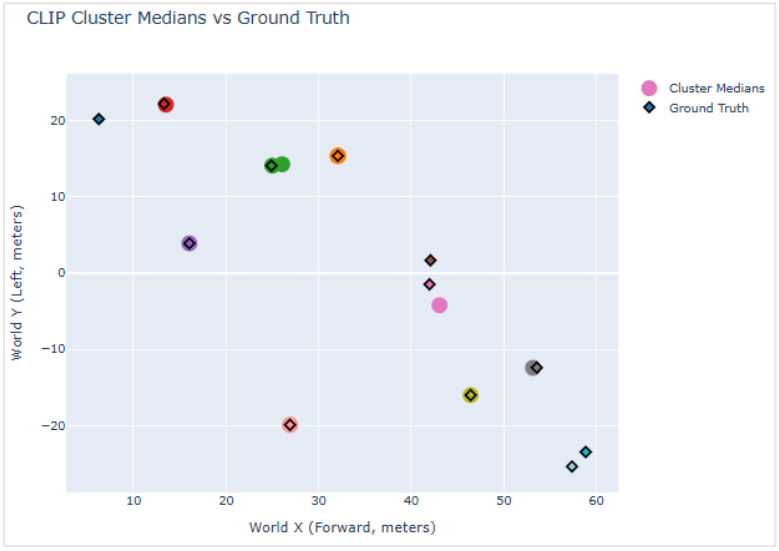

Figure 7. (CLIP): The most refined clustering, combining geometric and semantic features, closely aligns with ground truth object positions; thus, achieving the best localization accuracy.

Clustering serves as a critical postprocessing step that transforms low-level, frame-by-frame gaze–object predictions into coherent object-level representations. In all three pipelines, DBSCAN groups predictions that are likely to belong to the same object instance across time and space. This grouping enables the computation of cluster representatives—typically, median or mean 3D coordinates—which serve as robust and interpretable estimates of the actual object locations (Table 2).

Table 2. Error after clustering for all three models.

Model	MAE (m)	RMSE (m)	Std. error (m)	Median error (m)	N points
M1-RG	4.402	6.200	4.366	2.662	3/19
M2-RF	2.943	4.014	2.729	2.861	6/20
M3-CLIP	0.408	0.930	0.835	0.000	10/12

The raw pipeline uses clustering directly on geometric ray–object intersections, the RF pipeline adds learned correction from the model, and the CLIP pipeline further incorporates semantic understanding from visual features. This progressive refinement allows for consistent processing and comparison across the different modeling approaches.

Each pipeline follows a modular workflow: data combination across multiple runs, model inference (if applicable), feature extraction and merging (CLIP embeddings, geometric features), DBSCAN clustering using the appropriate feature space and similarity metric, cluster representative extraction (median/mean), and visualization and ground truth comparison. This consistent architecture allows for systematic evaluation of each pipeline's performance and a clear understanding of how geometric, learned, and semantic information contribute to gaze–object localization. Clustering acts as the unifying step that enables object-level reasoning and evaluation across all approaches.

4. Discussion

The primary goal of this effort was to develop a data-driven fusion pipeline where we found that leveraging CLIP dramatically improves 3D gaze–object localization over raw geometric projection alone. The Raw Geometric baseline produced MAEs of 2.512 and 3.595 m in Runs 1 and 2, respectively, reflecting the limitations of uncorrected gaze and camera geometry. Introducing a Classic Model trained solely on 18 geometric features reduced MAE to 0.350 and 0.648 m, confirming that ML can effectively compensate for systematic sensor biases and drift.

Building on this strong baseline, the CLIP-Augmented Model combined the same geometric inputs with four principal components derived from CLIP embeddings.

In Run 1, this resulted in a further MAE reduction from 0.350 to 0.343 m (2.0% improvement) and an RMSE improvement from 0.972 to 0.961 m (1.13% improvement). In Run 2, the MAE dropped from 0.648 to 0.551 m (14.9% improvement)—although the RMSE increased slightly from 1.296 to 1.348 m (3.97% decline). These findings suggest that CLIP embeddings provide complementary semantic cues that can enhance prediction robustness, particularly by reducing large-error outliers.

A notable feature of both the Classic and CLIP-Augmented Models is the zero median error, indicating that over half of all gaze–object predictions align exactly with ground truth. While this high precision underscores model effectiveness, it also highlights the importance of examining outlier behavior; the CLIP branch's contribution is especially evident in Run 2, where it sharply reduced MAE, showing potential for further gains through improved embedding integration and model tuning.

The adoption of the STAIT data format was instrumental in achieving these results. By standardizing coordinate frames and data schemas across HoloLens sensor streams—including eye tracking, inertial measurement units, head pose, and video—STAIT ensures that geometric and visual features align consistently in space and time; thus, enabling repeatable feature extraction and model training in multiagent mixed-reality experiments.

Future work should pursue refined fusion strategies such as attention-based gating or learned late-fusion networks to dynamically weight geometric versus visual inputs on a per-frame basis. Expanding the dataset to include additional participants, varied environments, and operationally relevant scenarios will further validate generalization. Finally, real-time deployment and evaluation in live field exercises will be critical to confirm the practical utility of semantically enriched gaze–object localization for human–machine teaming.

5. Limitations

The current 3D gaze–object localization pipeline assumes objects are grounded on a flat plane and uses simple bounding box logic to associate gaze points with detected objects. This leads to two key limitations:

1) Stacked Object Ambiguity: When multiple objects are vertically aligned (i.e., a box on a table), the system only considers the first bounding box containing the gaze point, often selecting the lower object and ignoring the one on top.

2) Objects in the Air: The 3D intersection logic projects gaze rays to a fixed ground height, making them incapable of correctly localizing objects that are suspended or elevated (i.e., hanging lights, drones, or shelves).

Due to the lack of height-based filtering, depth prioritization, and occlusion handling, the system cannot reliably distinguish between objects at different elevations, leading to incorrect associations in multilevel or cluttered scenes.

6. Conclusion

The goal of this project was to demonstrate the feasibility of accurate 3D gaze–object localization in mixed-reality environments using data collected from HoloLens headsets and processed through a standardized, multi-agent spatiotemporal framework (i.e., STAIT). To this end, we developed and evaluated three pipelines: raw geometric projection, a Classic Model based on geometric feature fusion, and a CLIP-Augmented Model that integrates distilled visual embeddings.

Using controlled field and mixed-reality experiments, the Classic Model reduced MAE from 2.512/3.595 m (raw geometry) to 0.350/0.648 m (Runs 1 and 2). Building on this, the CLIP-Augmented Model, by concatenating four PCA-derived CLIP components with the same geometric features, achieved an MAE of 0.343 m and RMSE of 0.961 m in Run 1, outperforming the Classic Model's MAE of 0.350 m and RMSE of 0.972 m. In Run 2, the CLIP model showed an even larger MAE improvement from 0.648 to 0.551 m, representing a 14.9% reduction; however, RMSE slightly increased from 1.296 to 1.348 m. Overall, this demonstrates that semantic context from visual embeddings provides complementary information to geometric cues, tightening the error distribution's tails and reducing large-error outliers, especially in more variable conditions.

These results validate that data-driven fusion of geometric and visual features can push localization accuracy beyond what either modality alone can achieve. The zero median error in all models indicates exceptionally high precision for the majority of predictions, while the lowered RMSE demonstrates enhanced robustness against challenging gaze–scene geometries.

We conclude that accurate and adaptive gaze–object localization is possible using today's off-the-shelf headsets, combined with a modular processing pipeline built on STAIT. While the improvements from adding CLIP features were small, they were consistent, showing that visual context can help. This sets the stage for future work in combining different types of data more effectively such as through attention mechanisms or late-fusion methods and brings us closer to building smart responsive mixed-reality systems for real-world training and mission use.

7. References

Beyerer J, Fuchs S, Arens M. Multimodal gaze estimation in tactical environments. IEEE Trans Hum-Machine Syst. 2021;51(3):234–245. https://ieeexplore.ieee.org/document/9445407

Kellnhofer P, Recasens A, Stent S, Matusik W, Torralba A. Gaze360: physically unconstrained gaze estimation in the wild. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 6912–6921. https://openaccess.thecvf.com/content_ICCV_2019/papers/Kellnhofer_Gaze360_Physically_Unconstrained_Gaze_Estimation_in_the_Wild_ICCV_2019_paper.pdf

Mora KAF, Odobez J-M. Geometric generative gaze estimation (G3E) for remote RGB-D cameras. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 1773–1780. https://openaccess.thecvf.com/content_cvpr_2014/papers/Mora_Geometric_Generative_Gaze_2014_CVPR_paper.pdf

Parnerkar P. CV0EyeTracking. Github.com; 2024 Aug 6. https://github.com/parnerka/CV-EyeTracking

Zemblys R, Niehorster DC, Komogortsev O, Holmqvist K. Using machine learning to detect events in eye-tracking data. Behav Res Meth. 2019;51(1):160–181. https://doi.org/10.3758/s13428-017-0860-3

Zhao Z-Q, Zheng P, Xu S-T, Wu X. Object detection with deep learning: a review. IEEE Trans Neural Networks Learn Syst. 2019;30(11):3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

List of Symbols, Abbreviations, and Acronyms

2/3D	two-/three-dimensional
AR	augmented reality
CLIP	Contrastive Language-Image Pretraining
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ID	identification
M1-RG	Model 1: Raw Geometric
M2-CM	Model 2: Classic Model
M3-CLIP	Model 3: CLIP-Augmented Model
MAE	Mean Absolute Error
ML	machine learning
PCA	Principal Component Analysis
RF	radio frequency
RGB	red, green, and blue
RMSE	Root Mean Squared Error
STAIT	Spatiotemporal Analysis and Inference Toolkit
UAS	unmanned aerial systems
VR	virtual reality
XR	extended reality

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB