EgoExOR: An Ego-Exo-Centric Operating Room Dataset for Surgical Activity Understanding

Ege Özsoy[*]
Technical University of Munich
Munich Center for Machine Learning
ege.oezsoy@tum.de

Arda Mamur[*]
Technical University of Munich
arda.mamur@tum.de

Felix Tristram
Technical University of Munich
Munich Center for Machine Learning
felix.tristram@tum.de

Chantal Pellegrini
Technical University of Munich
Munich Center for Machine Learning
chantal.pellegrini@tum.de

Magdalena Wysocki
Technical University of Munich
Munich Center for Machine Learning
magdalena.wysocki@tum.de

Benjamin Busam
Technical University of Munich
Munich Center for Machine Learning
benjamin.busam@tum.de

Nassir Navab
Technical University of Munich
Munich Center for Machine Learning
nassir.navab@tum.de

Abstract

Operating rooms (ORs) demand precise coordination among surgeons, nurses, and equipment in a fast-paced, occlusion-heavy environment, necessitating advanced perception models to enhance safety and efficiency. Existing datasets either provide partial egocentric views or sparse exocentric multi-view context, but do not explore the comprehensive combination of both. We introduce EgoExOR, the first OR dataset and accompanying benchmark to fuse first-person and third-person perspectives. Spanning 94 minutes (84,553 frames at 15 FPS) of two emulated spine procedures, Ultrasound-Guided Needle Insertion and Minimally Invasive Spine Surgery, EgoExOR integrates egocentric data (RGB, gaze, hand tracking, audio) from wearable glasses, exocentric RGB and depth from RGB-D cameras, and ultrasound imagery. Its detailed scene graph annotations, covering 36 entities and 22 relations (568,235 triplets), enable robust modeling of clinical interactions, supporting tasks like action recognition and human-centric perception. We evaluate the surgical scene graph generation performance of two adapted state-of-the-art models and offer a new baseline that explicitly leverages EgoExOR's multimodal and multi-perspective signals. This new dataset and benchmark set a new foundation for OR perception, offering a rich, multimodal resource for next-generation clinical perception. Our code and data are available at https://github.com/ardamamur/EgoExOR.

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

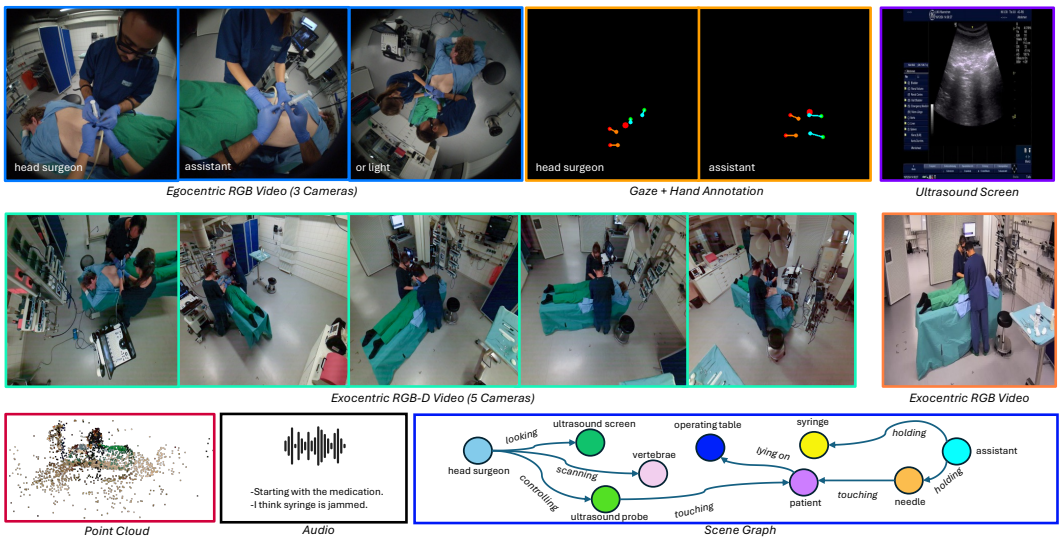

Figure 1: Overview of one timepoint from the EgoExoR dataset, showcasing synchronized multi-view egocentric RGB and exocentric RGB-D video streams, live ultrasound monitor feed, audio, a fused 3D point-cloud reconstruction, and gaze, hand-pose and scene graph annotations. The "closeTo" predicate is not visualized for brevity.

Table 1: Comparison of EgoExOR with existing operating room (OR) datasets. EgoExOR is the first to combine synchronized multi-view egocentric and exocentric recordings with gaze, hand pose, screen capture, and dense scene graph annotations.

Dataset	Ego.	Multi-View Ego	Multi-View Exo	Gaze	Hand Pose	Screen Recording	Scene Graphs	Annotated Timepoints
MVOR [1]			✓					732
4D-OR [2]			✓				✓	6,743
MM-OR [3]			✓			✓	✓	25,277
EgoSurgery [4, 5]	✓			✓				15,437
EgoExOR (Ours)	✓	✓	✓	✓	✓	✓	✓	84,553

1 Introduction

Modern operating rooms (ORs) are dynamic, safety-critical environments where diverse agents, such as surgeons, nurses, anaesthetists, mobile robots, and the patient, must coordinate seamlessly within a tightly confined space [6, 7]. Every participant has a distinct role, visual perspective, and attentional focus shaped by their responsibilities: a nurse may monitor tool availability, a surgeon may focus on a sub-millimetre needle movement, while an anaesthetist may track patient vitals. Crucially, a scene can evolve in seconds from a calm setup to a crowded, high-stakes intervention. This dynamic development and the multiplicity of perspectives makes the OR a uniquely complex environment, where a momentary lapse or misjudgment can jeopardize patient safety.

Capturing and understanding this complexity requires perception models that move beyond static, ceiling-mounted views and integrate multi-perspective, task-driven viewpoints that reflect the rich structure of human attention and interaction. Perspective-aware and accurate 4D scene understanding, capturing spatial and temporal dynamics of surgical workflows, would enable applications such as automatic documentation, improved team coordination, context-aware robotic or AR guidance, paving the way toward surgical automation. However, enabling such capabilities and making the first steps towards automated surgery is first and foremost **a data problem**. Progress across computer vision has consistently followed the release of large-scale public datasets, such as MNIST [8] and ImageNet [9] in early classification; KITTI [10], Cityscapes [11], and Waymo Open [12] for autonomous driving and Epic-Kitchens [13], Ego4D [14], and Ego-Exo4D [15] for egocentric activity understanding. Surgical Data Science (SDS) has seen similar advances in endoscopic vision, supported by internal

datasets for laparoscopy and arthroscopy [16, 17]. Existing OR datasets, such as MVOR [1], 4D-OR [2], MM-OR [3], and EgoSurgery [4, 5], have advanced scene modeling through multi-view, multimodal, or egocentric data; however, they lack the critical combination of multi-perspective egocentric and exocentric views, gaze, hand pose, and dense scene graphs. This limits holistic OR perception, specifically where multiple agents interact, occlusions are common, and fine-grained manipulations must be inferred from both global context and individual perspectives, hindering applications like context-aware robotic guidance and real-time team coordination.

These limitations highlight the need for a paradigm shift, from relying solely on static, top-down exocentric views to including task-specific perspectives of the OR staff. These perspectives often face the sterile field, circumventing the frequent occlusions in ceiling mounted cameras. Further, each team member has a unique view of the procedure, showing detailed gestures, tools and anatomical landmarks. Modern smart-glasses allow to record sub-millimetre eye-tracking; gaze vectors and hand-tracking, additionally to egocentric views, where each actors' gaze reveals intent and focus and their hand movements reflect granular actions such as grasping a scalpel or manipulating a needle. Together, these elements provide a comprehensive view of surgical activities from each staff's standpoint, essential for developing advanced computer vision models tailored to the OR. Furthermore, as wearable devices do not require modifications to the OR environment, they can be integrated more easily into existing operating rooms.

No existing dataset provides multi-member egocentric perspectives synchronized with exocentric OR views and dense frame-level scene graph annotations. To this end, we introduce EgoExOR, a new dataset and benchmark designed to advance OR scene understanding through an egocentric lens. For the first time, EgoExOR combines multiview egocentric recordings from wearable glasses, including RGB video, audio, gaze, and hand pose data, with multiview exocentric views from RGB-D cameras and screen recordings of ultrasound imaging, as visualized in Figure 1 and described in Table 1. Recorded in a simulated OR environment to ensure ethical and practical feasibility, EgoExOR spans 94 minutes across 41 takes, comprising 84,553 timepoints recorded at 15 FPS. Simulating two key surgical workflows, needle insertion and microsurgery, it delivers synchronized, multimodal data accompanied by scene graph annotations. We establish a new benchmark for surgical scene graph generation, evaluate two state-of-the-art OR scene graph generation models, and introduce a new baseline that leverages all the modalities in EgoExOR.

By uniting the staff members' visual experience with multi-view context and structured annotations, EgoExOR sets the stage for methods that reason jointly about gaze, dexterous manipulation, and the wider clinical scene. Finally, as EgoExOR is the first dataset to combine multiple egocentric and multiple exocentric viewpoints in a synchronized setup and capture simultaneous parallel actions from different agents, we believe it will serve as a cornerstone for the next generation of OR perception models and, more broadly, for any domain where human expertise, attention, and fine motor skill intersect in complex, occluded environments.

2 Related Work

Surgical Data Science. Surgical Data Science (SDS) has advanced significantly, leveraging deep learning for tasks like action recognition [18, 19], phase identification [20], instrument detection [21], enabled by large datasets like Cholec80 [16] and ArthroPhase [17]. These datasets focus on internal patient views (e.g., laparoscopy, arthroscopy), offering rich annotations but missing the broader OR context, such as interactions among clinical staff or external equipment. Capturing the entire operating theatre is more difficult because cameras must be non-intrusive, patient privacy is paramount and lighting is challenging. The Multi-View Operating Room (MVOR) dataset [1] uses three synchronized RGB-D cameras, capturing 732 frames with coarse pose labels, but its limited scale and coarse annotations restrict its utility. 4D-OR[2] introduced semantic scene graphs for holistic OR modeling, annotating 6,743 timepoints from six ceiling-mounted RGB-D cameras with clinical roles, tools, and interactions. MM-OR [3] scaled this effort, integrating multimodal data and panoptic segmentations for robotic knee surgeries, with an order of magnitude more annotations. However, both rely solely on exocentric views, which do not capture the surgical team members' unique perspectives, are susceptible to occlusions, and lack gaze or hand pose data. These limitations reduce their utility to mainly analysing the top-down perspective on the unobstructed OR room.

Egocentric Vision. In the wider field of Computer Vision multiple ego- (and exo-)centric datasets have been proposed to tackle video understanding tasks. Epic-Kitchens [13] for example captures 100 hours of cooking tasks with annotations for actions, objects, and narrations. Epic-Fields [22] extends Epic-Kitchens to include camera poses and sparse SfM pointclouds, enabling object tracking and more comprehensive 3D understanding. HD-EPIC [23] (41h) extends 3D reconstruction annotations and manually curate coarse meshes of the entire scene, which is highly promising for improving the precision of object tracking methods. All the aforementioned datasets, however, focus on cooking and are recorded solely in kitchens, limiting their broader applicability. In contrast, Ego4D [14] was recorded in diverse scenes and settings and spans 3,670 hours, introducing a variety of different benchmarks, from episodic memory over hand-object interaction to action anticipation. For multimodal sensing, Ego-Exo4D [15] combines synchronized egocentric and exocentric views across a combined 1,286 hours (221.26 ego-hours) of skilled activities (e.g., sports, cooking, music), enabling 3D human pose reconstruction and skill assessment. Another multimodal dataset was proposed in HOI4D [24], where a head-mounted RGB-D camera was used to capture egocentric video of human-object interactions. Aria Digital Twin (ADT) [25], recorded with Project Aria glasses, provides 3D reconstructions, audio, and gaze. For procedural tasks, Assembly101 [26] offers 4,321 clips captured from ego and exo perspectives of toy assembly with hand pose annotations.

While these datasets provide important building blocks for general computer vision tasks such as human-object interaction and video understanding in diverse settings, methods developed on them will not generalise to the OR without training data and method design that takes into account the complicated environment of the OR [3]. Except for one work, EgoSurgery [4], there have been no significant efforts to capture egocentric surgical videos, which could provide crucial views of the surgical staffs hands. Specifically, EgoSurgery provides a first-person view from surgeon-mounted camera, with phase and tool labels, capturing actions like suturing but it lacks perspectives from the other staff, exocentric context, team dynamics, or dense scene graphs.

In summary, no single dataset integrates synchronized egocentric and exocentric video, eye tracking, 6-DoF hand-tool trajectories, and dense scene graphs in a clinically realistic OR, which is limiting development of holistic perception methods. EgoExOR addresses this gap with a multimodal, multi-view dataset tailored to precision tasks like needle insertion and microscopic operations, featuring rich annotations that enable reasoning about surgeon attention, intent, action, and clinical context, paving the way for next-generation ego-exocentric OR perception.

3 Dataset Acquisition

This section details the acquisition of the EgoExOR dataset, including the recording environment, sensor configurations, participant roles, and emulated surgical procedures, outlining the methodology for acquiring multimodal, multiview data in an OR setting.

3.1 Recording Environment

EgoExOR was recorded in an university-affiliated surgical simulation center, previously utilized for the 4D-OR [2] and MM-OR [3] datasets. The center is equipped with surgical tables, anesthesia machines, overhead surgical lights, and standard OR equipment. The layout includes a sterile field with instrument tables and a circulation area for surgical team movement, ensuring a controlled setting for high-fidelity surgical simulations approximating real-world conditions *without* involving real patients. This approach mitigates privacy and safety concerns associated with patient data. While recording data from live surgeries would provide the highest level of realism, such recordings are rarely made publicly available due to strict privacy regulations and ethical considerations [27, 28]. Simulation-based data collection thus remains the most practical approach for creating open, reproducible datasets for surgical scene understanding.

3.2 Technical Setup

To capture the rich dynamics of each scenario, we instrumented the environment and participants with a comprehensive set of synchronized sensors, enabling both egocentric (first-person) and exocentric (third-person) views of the action. The following describes the equipment, camera placements, synchronization, and data processing pipeline.

Egocentric Recording. We used Project Aria Glasses [29] to capture first-person perspectives from participants (head surgeon, assistant, circulating nurse (circulator), anaesthetist) and non-human viewpoints (surgical microscope for MISS, OR light for Ultrasound). Key specifications were:

- **RGB Cameras**: 1440×1440 resolution at 15 FPS, providing first perspective views of the wearer, essential for resolving subtle actions like needle handling or micro-suture pickup.
- **Eye Tracking Cameras**: 320×240 resolution delivers gaze samples at 120 Hz with sub-millimeter accuracy (2D pixel + depth), enabling precise analysis of surgical intent (e.g., the surgeon's gaze switching between the ultrasound screen and patient).
- **Microphones**: Stereo at 48 kHz, capturing dialogue and ambient sounds.

Wearable cameras were also placed on the microscope (MISS) and OR light to simulate views that would be possible to get from next-generation robotic equipment or OR lights equipped with cameras.

Exocentric Recording. To capture the global OR context, we used Azure Kinect Cameras, strategically positioned at the ceiling for full-room coverage. Each records RGB-D images at 15 FPS, offering complementary external views of staff interactions and equipment, and enabling the creation of colored point clouds per timepoint. Camera placements were calibrated using a checkerboard pattern to compute intrinsic and extrinsic parameters, enabling spatial alignment and 3D reconstruction.

Ultrasound Screen Recordings. We used an HDMI capture device to record the screen of the ultrasound machine, providing real-time imaging at 15 FPS.

Synchronization. Azure Kinect cameras were connected in a master-slave configuration, ensuring frame-level alignment. To ensure all wearable cameras, ultrasound recording, and external recordings were in sync with each other, we used a clapper at the beginning of each take, clearly visible to every camera and audible to all Aria microphones, and manually synchronized the streams. We did not observe any drift in the streams during the duration recordings.

3.3 Participants and Roles

EgoExOR features emulated surgical teams composed of biomedical engineers trained to enact specific clinical roles. To enhance procedural variability and mitigate role-specific biases, roles were rotated across recordings, with individuals performing multiple roles and each role executed by different participants. All participants provided written consent for the recording and public release of the dataset, ensuring compliance with ethical standards.

- **Head Surgeon**: Directed the procedures, simulating primary tasks like needle insertion or disc removal.
- **Assistant**: Supported the head surgeon by handing instruments, adjusting equipment, and monitoring the surgical field.
- **Circulator**: Managed the OR environment, prepared tools, and maintained the sterile field.
- **Anaesthetist**: Oversaw the patient's vitals and anesthesia.

3.4 Surgical Procedures

EgoExOR models two precision-oriented interventions, **Ultrasound-Guided Needle Insertion (UI)** and **Minimally Invasive Spine Surgery (MISS)**, selected for their prominence in modern spine practice, their representation of distinct imaging-guidance paradigms and their sequential role in treating lumbar disc herniation, a leading cause of lumbar spine surgeries affecting approximately 5 to 20 per 1,000 adults each year [30] and causing radicular pain or lower back pain due to disc pressure on spinal nerves. Lumbar injections such as UI are a widely practiced intervention, with over 1 million lumbar epidural steroid injections administered annually in the United States alone [31]. MISS, particularly lumbar microdiscectomy, is the next step when the injections prove insufficient, removing the herniated disc fragments through a minimally invasive approach, with more than 300,000 procedures performed annually in the U.S. [32]. Although UI and MISS differ in invasiveness and therapeutic scope, they both heavily rely on precise, skillful interaction between the clinicians, tools and patient anatomy in a complex OR environment, making them suitable for studying OR dynamics in a clinically meaningful context. EgoExOR emulated these procedures

following surgical textbooks [33], training videos [34–36], and consultation with clinicians, ensuring alignment with clinical practice guidelines. We defined a phase taxonomy based on standard clinical workflows, segmenting procedures into distinct steps. Recording was organized around this phase structure, and each take corresponds to one or more procedural phases. All takes were scripted and rehearsed in advance to ensure both authenticity and clinical relevance.

Ultrasound-Guided Needle Insertion (UI). Simulates an epidural steroid or nerve-root block, where a spinal needle is advanced under real-time ultrasound. Phases:

- **Patient Introduction & Positioning**: The circulator prepares the OR, verbally confirming tool readiness (e.g., "Syringes ready, probe cover in place"), while the head surgeon and assistant position the patient and apply antiseptic.
- **Ultrasound Setup & Target Identification**: The head surgeon adjusts the ultrasound probe to locate landmarks, with the assistant preparing the needle and confirming angles.
- **Medication Injection**: The head surgeon inserts the needle, injects steroids, and observes the spread on the ultrasound screen, while the circulator monitors vitals.
- **Post-Procedure Cleanup**: The needle is removed, a dressing applied, and the team cleans the probe and OR.

Minimally Invasive Spine Surgery (MISS). Simulated using a surgical microscope and tubular retractors to remove herniated disc fragments. Phases:

- **OR Preparation & Patient Setup**: The circulator organizes instruments and checks the microscope, while the anaesthetist sets up monitors. The assistant drapes the patient, and the head surgeon confirms positioning.
- **Incision & Initial Access**: The head surgeon makes a small incision, places dilators, and uses the microscope to target the disc, with the assistant monitoring fluoroscopy.
- **Microscope-Guided Discectomy**: The head surgeon removes disc fragments using micro-forceps, confirming decompression, while the assistant ensures a clear field.
- **Wound Closure & Turnover**: The incision is closed with sutures, the patient is woken, and the team sterilizes equipment.

Furthermore, we also designed realistic deviations from typical protocols to reflect the variability observed in surgeries. We scripted complication for each phase, such as dropped instruments (e.g., assistant dropping gauze), needle contamination, ultrasound gel drying, microscope misalignment, syringe jams, or vital-sign fluctuations. Participants responded per standard protocols (e.g., replacing contaminated tools, re-sterilizing gloves), increasing diversity and creating challenging edge cases.

4 Dataset Description

This section describes the post-acquisition pipeline that transforms raw multimodal recordings into a structured dataset tailored for OR perception research. We outline the data modalities, processing steps, recording segmentation, statistical overview, and annotation process, providing a comprehensive view of EgoExOR's composition and utility.

4.1 Data Processing and Modalities

The EgoExOR dataset comprises multiple synchronized data modalities captured in an OR environment. After acquisition, all data streams undergo a standardized processing pipeline to produce a final, compact dataset suitable for efficient training and evaluation. The following summarizes all included data modalities, their processing and final representations as available in the dataset:

- **Egocentric RGB Video**: Multi-view RGB, down-sampled to 336×336 pixels, providing first-person perspectives of the OR.
- **Exocentric RGB-D Video**: Multi-view RGB and depth data, downsampled to 336×336 pixels, offering external perspectives of the OR scene.

Table 2: Distribution of surgical takes, duration, and frames across procedures in EgoExOR.

Procedure	# Takes	Time [min]	# Frames
Ultrasound-Guided Injection (UI)	24	69	62,547
Minimally Invasive Spine Surgery (MISS)	17	25	22,006
Total	41	94	84,553

- **Point Cloud**: Depth maps are processed into colored point clouds using calibrated camera extrinsics and reduced to 2,500 points per timepoint.
- **Gaze Tracking**: Eye-tracking data provided as normalized 2D coordinates relative to the egocentric image frame, combined with depth estimates for each gaze sample.
- **Hand Tracking**: Hand pose data from the glasses, with 16 points tracked (8 for the left wrist and palm, 8 for the right wrist and palm).
- **Audio**: Two-channel stereo audio at 48 kHz, normalized to a [-1, 1] range, capturing verbal communication and environmental sounds. Both full audio recordings, aligned with the take, and one-second audio snippets, aligned with individual timepoints are provided.
- **Ultrasound Screen Recordings**: Video captures of ultrasound display, downsampled to 336×336 pixels, enabling synchronized analysis alongside egocentric and exocentric views.

All modalities are temporally synchronized at a uniform rate of 15 FPS, and stored in a consolidated HDF5 archive for easy access. This ensures that visual, auditory, spatial, and other signals are consistently aligned across all timepoints, enabling robust multimodal learning and evaluation.

4.2 Dataset Composition and Statistics

EgoExOR consists of 41 takes capturing clinically inspired phases of two emulated surgical procedures: Ultrasound-Guided Injection (UI) and Minimally Invasive Spine Surgery (MISS), as detailed in Table 2. All takes are annotated and synchronized across modalities. The dataset contains a total of 84,553 frames recorded at 15 FPS, spanning 94 minutes. For benchmarking, the dataset is split into training, validation, and test sets at take level, ensuring phase and scenario diversity. Specifically, 26 takes are used for training, 8 takes for validation, and 7 takes for testing. The dataset is 195 GB in size, providing a comprehensive resource for developing and benchmarking OR perception models. A detailed description of the data structure and modalities is provided in the supplementary material.

Edge-Case Deviations. In addition to "normal" workflows, EgoExOR systematically integrates edge cases that represent realistic deviations from surgical protocols (e.g., dropped instruments, gel drying, microscope misalignment). Out of the 41 total takes, 15 contain no scripted deviations, while 26 takes include at least one edge case. These 26 takes account for approximately 54k out of 84k total frames (∼64% of the dataset). Importantly, edge cases are restricted to specific sub-phases, and not all frames within such takes depict anomalies. Each edge-case take typically contains multiple instances of the targeted deviation, and role assignments are varied across takes to capture different team dynamics. To support systematic evaluation, we provide a metadata file mapping each take to its surgical sub-phase and indicating whether it includes edge-case events.

4.3 Annotations

EgoExOR follows previous works [2, 3], and includes per-frame scene graph annotations to provide structured semantic context, capturing entities (e.g., surgeons, tools, patient) and their relations (e.g., inserting, cutting). Annotations were created manually using a custom interface that allowed annotators to navigate frames, zoom into details, and overlay multimodal data (e.g., gaze points on RGB frames). For each frame, one trained annotator labeled entities and relations, with a second annotator verifying the output to ensure accuracy.

The annotation schema comprises **36 entity classes** (e.g., `head_surgeon`, `scalpel`, `ultrasound_probe`, `patient`) and **22 relation classes** (e.g., `holding`, `using`, `closeTo`), reflecting clinical roles, tools, anatomical landmarks, and their interactions. On average, each frame contains **6.8 ± 2.5 relation triplets** (median = 7, max = 13), with a total of 568,235 triplets across all 84,553 frames. In the appendix we provide exact statistics for the distribution of all classes.

5 Scene Graph Generation Benchmark

This section presents our benchmark for surgical scene graph generation, trained and evaluated on the EgoExOR dataset. All experiments were conducted on a single NVIDIA A40 GPU (48GB) using PyTorch 2.0.1, CUDA 11.7, and Python 3.11, and training time was approximately 5 days.

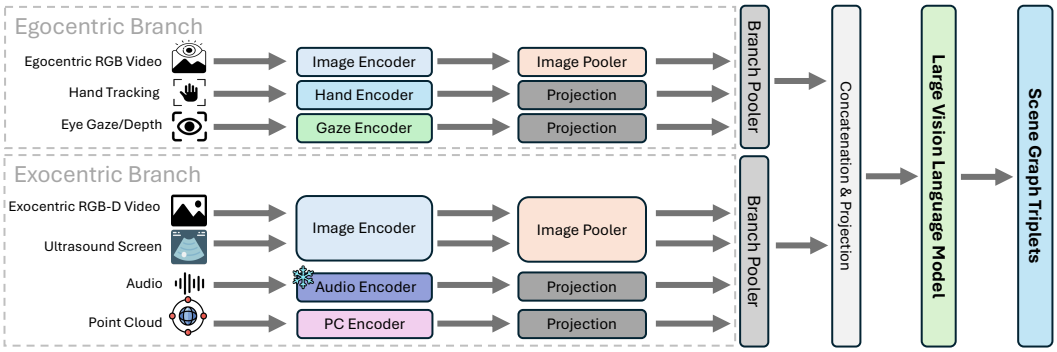

Figure 2: Overview of the proposed EgoExOR model for surgical scene graph generation. The model employs a dual-branch architecture to separately process egocentric and exocentric modalities. Fused embeddings are passed to a large language model (LLM) to autoregressively generate scene graph triplets representing entities and their interactions.

The surgical scene graph generation task [37, 2, 38, 39] aims at generating structured graph representations of OR scenes with nodes as entities (e.g., surgeon, patient, tools) and edges as interactions (e.g., *holding*, *lying on*), summarizing the semantics of the scene. Following prior work [40, 3], we represent scene graphs as triplets (*subject, predicate, object*), capturing dynamic interactions like *(assistant, aspirating, patient)* and spatial relationships like *(patient, lying on, operating table)*. We train and evaluate two baseline models, ORacle [40] and MM2SG [3], adapted to process EgoExOR's ego-exo images. ORacle employs a 2D multi-view visual encoder to embed OR scenes, where as MM2SG extends this incorporating additional modalities such as point clouds, ultrasound screen recordings, and audio. Both use a large language model (LLM) to autoregressively predict scene graph triplets. Both process egocentric and exocentric RGB inputs jointly within a single shared encoder, and lack dedicated fusion mechanisms for perspective-specific features, and are unable to leverage EgoExOR's hand and gaze tracking signals.

EgoExOR Model. To fully exploit EgoExOR's rich multi-perspective data, we introduce a new baseline model featuring a dual-branch architecture Figure 2. The *egocentric branch* processes first-person RGB, hand pose, and gaze data, while the *exocentric branch* handles third-person RGB-D, ultrasound recordings, audio, and point clouds. Each branch uses a 2-layer transformer to fuse its inputs into N feature embeddings. These are concatenated and fed into the LLM for triplet prediction. By explicitly separating and fusing perspective-specific features, our model better captures actions and staff interactions, outperforming single-stream baselines in modeling complex OR dynamics.

Evaluation Metric. Following established protocols [2, 40, 3], we evaluate performance using the macro F1-score over predicates, assigning equal weight to each class to account for class imbalance.

Implementation Details. All models use LLaVA-7B [41] as the starting point, with Vicuna-7B [42] as the LLM and CLIP ViT [43] as the image encoder. We fine-tune using LoRA [44] for the LLM and adapt the last 12 layers of the image encoder for the OR domain, following MM2SG [3]. The number of fixed image tokens N is set to 576. Audio is encoded with CLAP [45], and point clouds with Point Transformer V3 [46]. For the EgoExOR model, both hand pose data and gaze/gaze-depth data are encoded using MLPs, and then projected into a shared latent space as single tokens. All models were trained for 4 epochs. The adapted ORacle and MM2SG baselines also process EgoExOR's egocentric RGB alongside exocentric data, however in a unified stream, lacking the EgoExOR model's specialized branch for egocentric signals.

Results. Overall the results, shown in Table 3 and Figure 3, demonstrate that more modalities lead to improved results, and while the existing works perform satisfactorily, the dual-branch EgoExOR model achieves the highest macro F1. Several predicates such as *injecting, aspirating, holding,*

Table 3: Scene graph generation results on EgoExOR. The table reports macro F1 scores for two surgical procedures: Ultrasound-Guided Injection (UI) and Minimally Invasive Spine Surgery (MISS). Used modalities such as egocentric and exocentric RGB images (Images), ultrasound screen (Ultra.), audio, point cloud (PC), gaze and hand pose (Hand) are indicated with a checkmark.

Model	Images	Ultra.	Audio	PC	Gaze	Hand	UI	MISS	Overall
ORacle [40]	✓						0.65	0.66	0.63
MM2SG [3]	✓	✓	✓	✓			0.72	0.66	0.67
EgoExOR (Ours)	✓	✓	✓	✓	✓	✓	**0.79**	**0.68**	**0.72**

controlling in EgoExOR rely on understanding transient tool-hand trajectories, and fine-grained action cues. This emphasizes the importance of explicitly modeling multiple viewpoints and leveraging all available modalities to improve OR scene understanding. However, it still struggles with low-frequency predicates such as *cutting, positioning*. We provide detailed per-predicate performance metrics and further visualizations in the supplementary material.

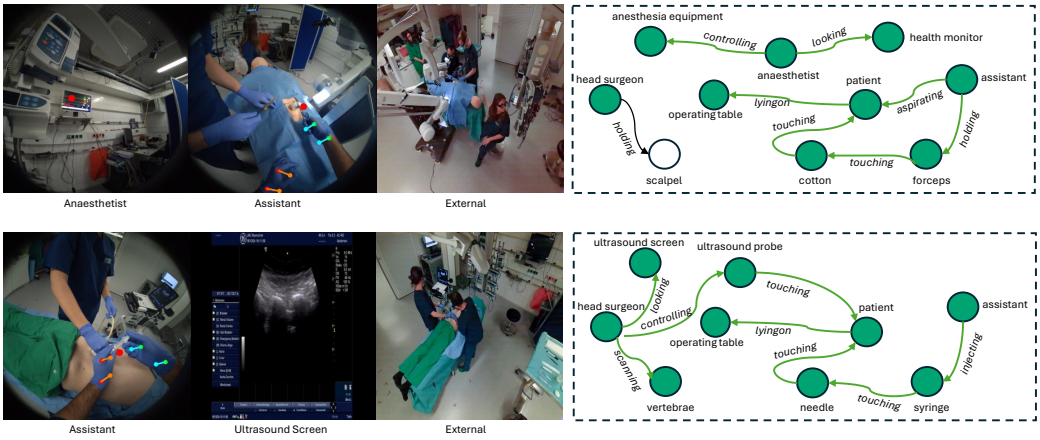

Figure 3: Qualitative examples from EgoExOR. Correctly predicted entities and predicates are highlighted in green, wrong ones are left white. The "closeTo" predicate is not visualized for brevity.

6 Conclusion

EgoExOR marks a significant advance in surgical data science, delivering a comprehensive dataset that captures the intricate dynamics of operating rooms through synchronized egocentric and exocentric viewpoints. Across 84,553 frames of two simulated spine procedure, it combines diverse modalities with rich scene graph annotations to model clinical workflows holistically. Our benchmark, comparing existing OR perception models against a new method designed for EgoExOR, underscores its value for developing intelligent systems, from automated documentation to context-aware robotic assistance. EgoExOR's fine-grained action detection, presents unique challenges, making it a valuable resource for advancing capabilities essential for advancing OR perception and human-centric AI in complex, safety-critical environments. We believe EgoExOR will catalyze innovation in OR perception, paving the way for smarter, safer surgical environments, with implications for any field requiring precise human-robot collaboration in complex environments.

Ethical Considerations. By employing simulated procedures, EgoExOR avoids privacy concerns inherent to clinical data, enabling unrestricted public release and reproducibility in alignment with NeurIPS ethical standards. All participants provided informed consent for data collection and publication. Potential positive societal impacts include accelerating the development of intelligent clinical assistance systems that improve surgical safety, support clinical decision-making, and enhance training through rich multimodal datasets. As with any technology in sensitive domains, there is also a potential negative impact if such models are deployed prematurely or without sufficient human oversight, potentially leading to over-reliance on automated systems in high-stakes clinical

environments. We emphasize that EgoExOR is intended as a research resource, not as a clinical decision-making system.

Limitations. While the simulated procedures in EgoExOR provide a valuable ressource for multi-perspective and multi-sensor OR understanding, its simulated setup can not fully reflect the intricacies of expert surgical performance in live clinical settings. Additionally, the recordings are limited to a specifically equipped operating room that enabled high-quality, synchronized multimodal data but may restrict the adaptability of models to varied OR configurations. Future work could expand this scope, building on EgoExOR's robust foundation.

Acknowledgements. We thank INM and Frieder Pankratz in helping setting up the acquisition environment, and Felix Holm and Miruna-Alexandra Gafencu for their help in data acquisition. Authors would like to thank Carl Zeiss AG for their partial support. We also thank Carl Zeiss AG for their partial financial support, and gratefully acknowledge Meta Reality Labs for providing Aria research glasses.

References

[1] Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. *arXiv preprint arXiv:1808.08180*, 2018.

[2] Ege Özsoy, Evin Pınar Örnek, Ulrich Eck, Tobias Czempiel, Federico Tombari, and Nassir Navab. 4d-or: Semantic scene graphs for or domain modeling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*. Springer, 2022.

[3] Ege Özsoy, Chantal Pellegrini, Tobias Czempiel, Felix Tristram, Kun Yuan, David Bani-Harouni, Ulrich Eck, Benjamin Busam, Matthias Keicher, and Nassir Navab. Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In *CVPR*, 2025.

[4] Ryo Fujii, Masashi Hatano, Hideo Saito, and Hiroki Kajita. Egosurgery-phase: A dataset of surgical phase recognition from egocentric open surgery videos. In *MICCAI*, 2024.

[5] Ryo Fujii, Hideo Saito, and Hiroki Kajita. Egosurgery-tool: A dataset of surgical tool and hand detection from egocentric open surgery videos. *arXiv preprint arXiv:2406.03095*, 2024.

[6] Florent Lalys and Pierre Jannin. Surgical process modelling: a review. *International Journal of Computer Assisted Radiology and Surgery, Springer Verlag*, 9:495–511, 2014.

[7] Lena Maier-Hein, Swaroop S. Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, Makoto Hashizume, Darko Katic, Hannes Kenngott, Michael Kranzfelder, Anand Malpani, Keno März, Thomas Neumuth, Nicolas Padoy, Carla Pugh, Nicolai Schoch, Danail Stoyanov, Russell Taylor, Martin Wagner, Gregory D. Hager, and Pierre Jannin. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, September 2017.

[8] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[15] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 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*, pages 19383–19400, 2024.

[16] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[17] Ali Bahari Malayeri, Matthias Seibold, Nicola Cavalcanti, Jonas Hein, Sascha Jecklin, Lazaros Vlachopoulos, Sandro Fucentese, Sandro Hodel, and Philipp Furnstahl. Arthrophase: A novel dataset and method for phase recognition in arthroscopic video. *arXiv preprint arXiv:2502.07431*, 2025.

[18] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 364–374. Springer, 2020.

[19] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78, 2022.

[20] Tobias Czempiel, Magdalini Paschali, Daniel Ostler, Seong Tae Kim, Benjamin Busam, and Nassir Navab. Opera: Attention-regularized transformers for surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 604–614. Springer, 2021.

[21] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 691–699. IEEE, 2018.

[22] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36:26485–26500, 2023.

[23] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. *arXiv preprint arXiv:2502.04144*, 2025.

[24] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[25] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023.

[26] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[27] Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Automatic operating room surgical activity recognition for robot-assisted surgery. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 385–395. Springer, 2020.

[28] Keqi Chen, Lilien Schewski, Vinkle Srivastav, Joël Lavanchy, Didier Mutter, Guido Beldi, Sandra Keller, and Nicolas Padoy. When do they stop?: A first step towards automatically identifying team communication in the operating room. *arXiv preprint arXiv:2502.08299*, 2025.

[29] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023.

[30] Kevin Y Heo, Janice M Bonsu, Sameer Khawaja, Anthony Karzon, Prashant V Rajan, Lauren A Barber, and Sangwook Tim Yoon. Database analysis comparing incidence and complication rates between inpatient and outpatient laminotomies for lumbar disc herniation. *North American Spine Society Journal (NASSJ)*, 18:100328, 2024.

[31] Judith A. Racoosin, Sally M. Seymour, Laurelle Cascio, and Rajdeep Gill. Serious neurologic events after epidural glucocorticoid injection—the fda's risk assessment. *New England Journal of Medicine*, 373(24):2299–2301, 2015.

[32] Chris D. Daly, Kai Zheong Lim, Jennifer Lewis, Kelly Saber, Mohammed Molla, Naor Bar-Zeev, and Tony Goldschlager. Lumbar microdiscectomy and post-operative activity restrictions: A protocol for a single blinded randomised controlled trial. *BMC Musculoskeletal Disorders*, 18:312, 2017.

[33] M. P. Steinmetz and E. C. Benzel, editors. *Benzel's Spine Surgery: Techniques, Complication Avoidance, and Management*. Elsevier, 5th edition, 2021.

[34] R3 Medical Training. Spine ultrasound injection course in action (888) 998-6343. YouTube, 2025. Available at `https://www.youtube.com/watch?v=icPj6_OTpEQ`, accessed May 11, 2025.

[35] Murat Karkucak. Ultrasound guided lumbar disc herniation injection. YouTube, 2022. Available at `https://www.youtube.com/watch?v=Vm6BTPdOw5A`, accessed May 11, 2025.

[36] Jon Kimball, Andrew Yew, and Daniel C. Lu. Minimally invasive surgery for lumbar microdiscectomy. YouTube, 2013. Available at `https://www.youtube.com/watch?v=aXyZ2FJMh2s`, accessed May 11, 2025.

[37] Mobarakol Islam, Lalithkumar Seenivasan, Lim Chwee Ming, and Hongliang Ren. Learning and reasoning with the graph structure representation in robotic surgery. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 627–636. Springer, 2020.

[38] Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1409–1417, 2024.

[39] Felix Holm, Ghazal Ghazaei, Tobias Czempiel, Ege Özsoy, Stefan Saur, and Nassir Navab. Dynamic scene graph representation for surgical video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 81–87, 2023.

[40] Ege Özsoy, Chantal Pellegrini, Matthias Keicher, and Nassir Navab. Oracle: Large vision-language models for knowledge-guided holistic or domain modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 455–465. Springer, 2024.

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[42] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

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

[44] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[45] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022.

[46] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state EgoExOR's contributions, including its novelty as the first dataset combining egocentric and exocentric views with scene graph annotations for OR scene understanding.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper highlights limitations in the corresponding section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper focuses on dataset creation and empirical contributions, not theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed recording setup, data processing, and annotation schema are provided, enabling reproduction of the dataset's multimodal streams and scene graphs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset is publicly released with Python/PyTorch loaders and instructions. Our code and data are available at `https://github.com/ardamamur/EgoExOR`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail this in the corresponding benchmark and experiments section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the evaluation standard of previous works.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the size of the dataset as well as the necessary compute required to train the models in section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Simulated data avoids privacy issues, and participant consent was obtained, aligning with NeurIPS ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the broader impacts section 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The simulated surgery dataset poses no high-risk misuse concerns, requiring no specific safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Not applicable as we collect our own data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: EgoExOR is released with comprehensive documentation, including metadata and usage instructions at `https://github.com/ardamamur/EgoExOR`.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve human research, as all participants are fellow researchers and biomedical engineers from the same chair, collaborating as part of shared research goals, and performing a simulated surgery without any risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subject research was conducted, as the dataset is simulated with participant consent for data use.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: While the main contribution, the dataset and benchmark, do not involve any LLMs, all the baseline models use an LLM in their architecture.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.