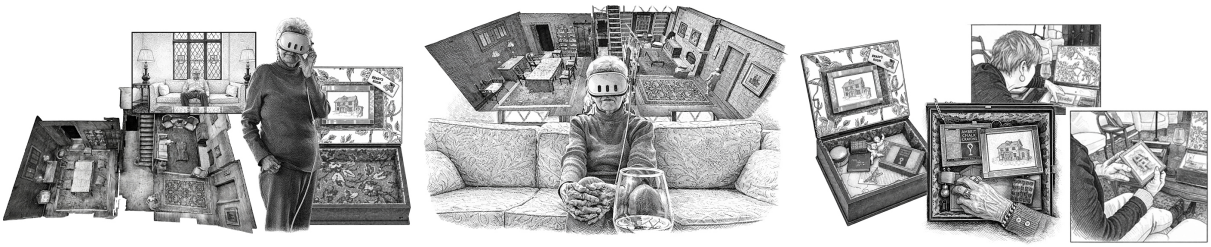

Figure 1. From left to right: images from the multisensory simulation study, including the environment, participants providing movement and autobiographical data, documentation of the sensory space, interactive tangible interfaces and demonstrative images

Multisensory Spatial Memory: A Large-scale, Synchronized Multi-Sensory Dataset of Embodied Interaction with Reconstructed Autobiographical Spaces

Anonymous CVPR submission

Paper ID *****

Abstract

Embodied intelligence requires models of how humans perceive and navigate space, yet most spatial datasets capture passive, single-modal data rather than active, perception-based interaction. We present the Jefferson Drive Simulation Dataset (JDSD), a large-scale multi-sensory corpus of data, capturing high-resolution 2D and 3D samples of a spatial environment, including tactile, olfactory, and interaction data, designed to capture the sensory experience of a specific space. For this dataset, we specifically chose to represent a memory-laden family home as a vehicle for creating a sensory-object-memory corpus of data.

Over the course of a year, we performed an exhaustive capture of a domestic environment in Pittsburgh PA, a single family home lived in by the same residents for over 50 years. By digitally scanning all of the available photographic data, analog video data, three dimensional maps, 360 photos of each room, as well as comprehensive data about the space, we reconstructed the environment at high fidelity. To add semantic and sensory content to the dataset, participants with experience of the space were instrumented with XR head-mounted displays, spatial pose trackers, and tangible haptic interfaces while exploring the multimodal dataset, including a Gaussian-splat reconstruction [25] of a residential home spanning ∼50 years of occupancy. During data collection, synchronized streams of egocentric video, 6-DOF pose, gaze, audio, and tangible object interaction were recorded and added as multi-sensory data to the digital dataset. Critically, each of these participants had extensive prior lived experience of the space, creating a unique setting where memory [5, 36], interaction, and perception are in dialogue with multimodal media.

Presented here, the dataset, totaling approximately 4TB across 800+ scanned objects and 18 rooms, provides a novel benchmark for studying multi-modal spatial understanding and embodied interaction with reconstructed environments, with applications in generalizable representations for manipulation in robotics and XR.

1. Introduction

Humans navigate space by both perceiving it and remembering it. Space stores experiences. When we use our visual navigation system, we overlay current sensory input with expectations built from prior experience, drawn from memory and emotional associations [5, 36]. However, dominant paradigms in spatial AI datasets prioritize the capture and processing of static environments, overlooking the richness

of how the human experience of space impacts our understanding of sensory intelligence.

This dataset emphasizes a modality that distinguishes human spatial behavior, i.e. memory-driven, anticipatory, affectively modulated navigation [42, 46]; an approach largely absent from training data. This work addresses that gap by introducing a dataset that combines 3D data, semantic data, and sensory data, specifically focused on a location and the memory of a group of individuals of the space.

We gathered data from human participants interacting with a photorealistic reconstruction of a domestic environment they had previously inhabited over a shared history spanning more than four decades. Using a combination of gaussian splatting [25], photogrammetry, world models, 3D scanning of 800+ individual objects across 18 rooms, and conversion of 40+ hours of 8mm/16mm film into spatial media, we reconstructed the environment, content, and sensory context at high fidelity in order to create a simulation space for capturing multisensory data triggered by associative memory [34].

Figure 4. Jefferson Drive Home in the 1980s

Participants were recruited from two closely related families with extensive prior experience in the chosen home. During the data gathering stage, the participants visited a physical facsimile of the home, complete with sensory details including furniture, food, and decorations. Participants navigated a virtual reconstruction of the original home while instrumented with multi-sensory capture hardware. The tracking data, sensory data, and semantic content of participants' memories were then added to the dataset.

The resulting dataset makes several novel contributions to multisensory AI. First, it captures human spatial interaction where memory and perception interact in measurable ways [5, 42], a condition unexplored in existing benchmarks [15, 16]. Second, our capture system synchronizes multi-modal streams including egocentric video, body and gesture poses, gaze, audio, and tangible haptic signals [32, 44]. Third, it includes ground-truth annotations linking spatial behaviors to autobiographical memory content [34, 46], enabling supervised learning of memory-driven spatial navigation.

Our results show that memory significantly modulates where humans look, reach, and focus [34, 46], emphasizing the need for novel sensory simulation training data for embodied agents operating in human environments. Additionally, the specific tactile qualities of materials in our study showed an impact on memory stimulation [31, 44]. Full results are pending and this paper illustrates only the dataset.

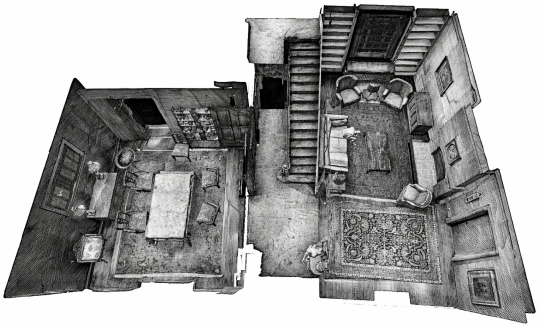

Figure 5. Drawing based on rendering of the first floor of the simulated space

2. Related Work

2.1. Multi-Sensory Spatial Datasets and Benchmarks

Existing large-scale spatial interaction datasets primarily capture passive scene observation [6, 9] or tactile-directed manipulation of isolated objects [3, 4, 45]. Cross-modal datasets combining egocentric video with gaze and pose are emerging [15, 16] but lack memory as an interaction modality. Lifelogging systems [18, 23, 27] represent an adjacent tradition of continuous personal data capture, but focus on indexing daily experience rather than modeling how memory shapes spatial behavior or manipulation. Our work, inspired by benchmarks in embodied AI [10, 37], contributes both real-human perceptual data and synthetic data gathered by simulated agents to create a novel multisensory multimodal dataset relevant to cross-modal spatial-temporal alignment and embodied manipulation benchmarking.

2.2. Human Spatial Memory and Navigation

Cognitive neuroscience has established that spatial memory and perception share neural substrates [5, 36] and that episodic memory significantly shapes navigational behavior [42, 43]. Rubin and Umanath [38] proposed a theory of event memory as a mental construction of a scene

Figure 2. Grid of archival living room scenes and objects arranged left-to-right, top-to-bottom: (a) Christmas tree with gifts and chair; (b) Coffee table with flowers and candles; (c) Piano with autumn decorations; (d) Mantel with framed portrait and floral arrangements; (e) Decorative eggs on mantel; (f) Window seat with dolls; (g) Piano with orange floral centerpiece; (h) Chair beside fireplace; (i) Living room with scattered holiday gifts; (j) Christmas display beneath tree; (k) Tree and gift bags near seating area; (l) Chair with wrapped gift; (m) Armchair with stuffed animals; (n) Fireplace with stockings and presents; (o) Chair with gift box; (p) Decorative tabletop arrangement; (q) Mantel with figurines and ornaments; (r) Close-up of figurines on piano; (s) Chair with birthday cushion; (t) Table setting with candles and flowers; (u) Holiday centerpiece with Santa figurine; (v) Living room with Christmas clutter; (w) Mantel with nativity-style figurines; (x) Chair with gifts and fireplace setting.

recalled as a single occurrence, relying on the hippocampus and visual stream as primary input. Behavioral studies have shown that humans exhibit anticipatory gaze toward semantically predicted locations [46] and preferentially encode emotionally significant spatial regions [34]. Research has also demonstrated that spatial memory is closely tied to physical place [14, 33], with location-based systems shown to enhance memory retrieval by anchoring experiences to their geographic origins. Our dataset operationalizes these findings at scale in a naturalistic XR setting, providing a large-scale behavioral corpus of spatial interaction based on memory and simulation, specifically with implications for embodied agents that must account for human spatial priors in domestic manipulation settings.

2.3. Emotion, Memory, and Sensory Activation

Memory is tightly interwoven with emotional and sensory experience, both of which modulate the spatial behaviors our dataset is designed to capture. Sas et al. [39] found that emotional arousal during encoding enhances the quality of episodic memory recall by over 50%, with direct implications for designing memory-triggered data collection protocols. Critically, for multi-sensory data collection, Amores et al. [13] demonstrated that olfactory wearables improve spatial navigation memory recall by easing cognitive load during encoding, inspiring our use of olfactory stimuli as

Figure 3. Equirectangular panoramas of a residential environment, ordered top-to-bottom and left-to-right. **First floor:** (a) entryway/hall, (b) dining room (c) breakfast room hallway/transition, (d) kitchen upstairs landing view, (e) living room, (f) den. **Basement:** (g) basement, pool room, **Second floor:** (g) stairwell landing, (h) hallway, (i) girl's bedroom, (j) parent's bedroom, (k) bathroom, (l) son's bedroom. **Basement/auxiliary:** (m) recreation/basement room, (n) storage/attic, (o) kitchen, (p) hallway.

a cross-modal (from physical to virtual) grounding mechanism during participant sessions.

2.4. 3D Reconstruction and Volumetric Capture for Embodied AI

Neural Radiance Fields [35] Gaussian splatting, and world models [25] have revolutionized photorealistic reconstruction of real environments, enabling high-fidelity scene representations suitable for embodied agent training [37]. Volumetric capture has been applied to preserve memories of real events [40] and enable shared multi-perspective playback in augmented reality [30]. Prior work on memory manipulation in XR has explored how reconstructed environments influence encoding and retrieval [2], while VR has been studied as a memory prosthesis for clinical applications [26]. Remixed Reality [29] demonstrated live 3D reconstruction of physical environments from external depth cameras for real-time spatial manipulation in mixed reality. Our pipeline extends these technologies by combining Gaussian splatting [25], world modeling, and object-level photogrammetry for 800+ individual items alongside multisensory stimuli, creating a hybrid representation space supporting both scene-level navigation and object-level interaction research directly relevant to the workshop's focus on generalizable multisensory models for manipulation.

2.5. Digital Twins and Spatial Simulation

Digital twins, virtual mirrors of physical environments supporting bidirectional physical-digital relationships, have emerged as a key paradigm for spatial simulation and embodied agent testing [17, 24]. Architectural digital twins have demonstrated value at urban scale [28, 47] and at the scale of individual interactive spaces [22]. Digital twin approaches have also been applied to human-robot interaction [12], enabling supervisors to control autonomous systems through immersive virtual representations of physical environments. Our dataset extends the digital twin paradigm to the multimodal, personal and autobiographical, creating a spatial memory corpus anchored in a single domestic environment documented across five decades, with significant implications for robot systems operating in human homes, specifically those of the elderly.

2.6. Tangible and Haptic Interfaces for Spatial Interaction

Prior work on tangible interfaces [20] and haptic feedback [32, 44] has shown that physical interfaces modulate spatial cognition and memory [31]. Csikszentmihalyi and Halton [8] established foundational theory on domestic objects as carriers of personal identity, directly relevant to our use of heirloom objects as memory stimuli. Bong et al. [1] found that tangible and nostalgic interface elements produce stronger engagement and retention than screen-based alternatives. Sion et al. [41] demonstrated that vibrotactile patterns mapped to episodic memories can physically instantiate past experiences, while Huber et al. [19] showed that tangible prototypes enhance reminiscence-based communication. Our dataset uniquely couples tangible physical objects configured as tactile memory boxes, containing original heirlooms, wallpaper samples, and olfactory cues,

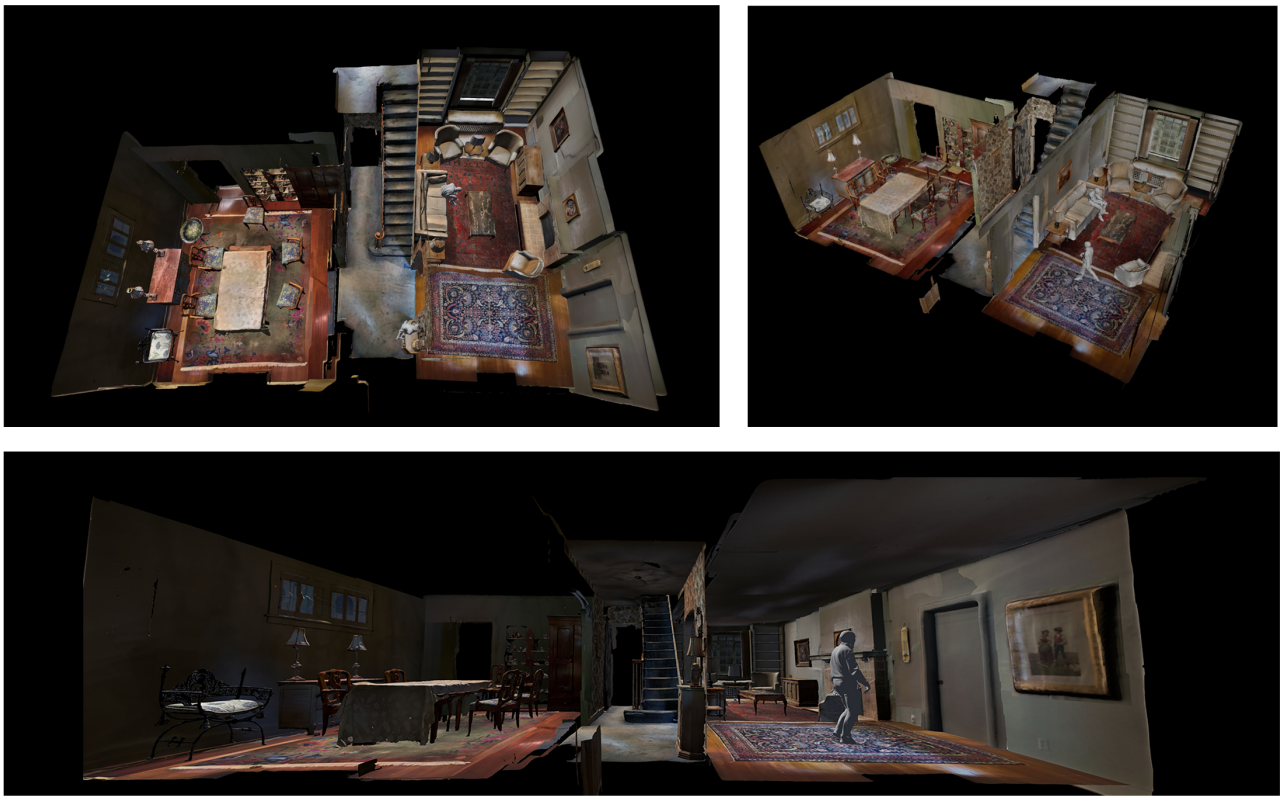

Figure 6. From left to right: 3D reconstructed environment from dataset from three perspectives, showing the egocentric point of view occupied by the subject

with perceptual data captured in a virtual reconstruction, enabling study of cross-modal embodied perception where physical and virtual information channels align.

3. The Jefferson Drive Spatial Memory Dataset

3.1. Environment Reconstruction

The reconstruction focused on a specific address on Jefferson Drive, a single-family residential home in Pittsburgh, PA, documented over the years 1965–2025 in various formats, such as celluloid film, analog photography, sound recordings, and high definition video. The physical space comprises 18 rooms across four floors. Our capture and reconstruction pipeline consisted of four complementary components.

Gaussian splat scene reconstruction for world models. Each room was photographed from 60-120 positions at multiple perspectives and processed using a structure-from-motion pipeline with both standard geometric photogrammetry and 3D Gaussian splatting [25] to produce photorealistic environment representations. Panorama photography was also captured as source data for the reconstruction. 360 panoramas (seen in Figure 3) were then used in combination with the hard surface photogrammetry alongside Gaussian splatting.

Object-level photogrammetry. Object-level photogrammetry was conducted for 800+ individual objects from the original collection of the family, including furniture, personal items, ephemera, decorative objects, and architectural details. Objects were scanned using a structured-light scanner and processed using a customized InstantMeshes [21]

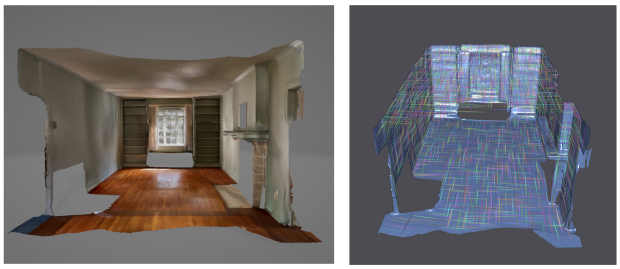

Figure 7. Retopologization process shown from the 3D scan showing the low poly instant meshes output

retopology workflow to produce clean meshes suitable for real-time rendering (process seen in Figure 9). Individual object scans were manually aligned and added to a dataset of individual elements configurable in room-specific datasets.

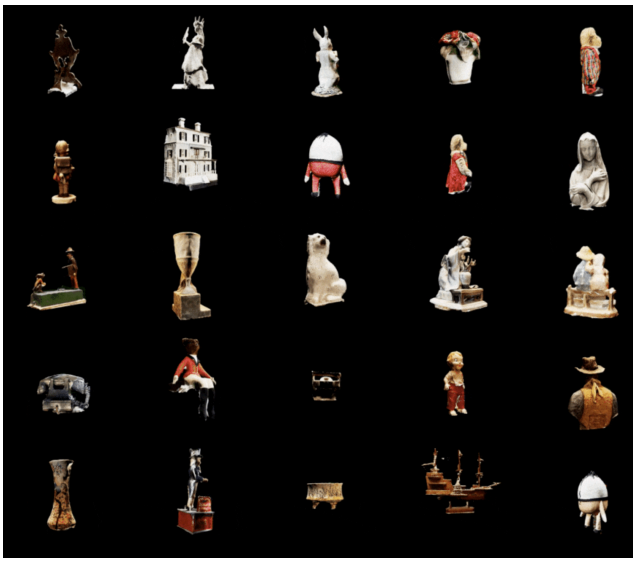

Figure 8. Rendered images from dataset capture showing the sensory integration of touch, sight, and memory during the data gathering process

Archival media integration. Over 40 hours of 8mm and 16mm home film were digitized and annotated with spatial metadata linking footage to specific rooms and time periods, comprising a spatially anchored dimension of the multimodal dataset. (process seen in Figure 9). The films were rendered as time based Gaussian splats [25], 4GDS movies, transposed into the original locations in the reconstruction.

Photo-based elements. Hundreds of family photographs were digitally scanned and processed with spatial metadata linking images to specific rooms and time periods. Using computer vision, room locations in photos were organized based on background content in each image.

Olfactory and haptic elements. Olfactory and physical sensory data was gathered through two tangible "memory boxes"—book-like wooden containers housing original wallpaper samples, carpet fragments, relevant fragrances, and personal artifacts from the home. These objects retained olfactory attributes of the original space and were used during semantic data gathering as a multimodal grounding stimulus [31, 44], creating associations of scents and textures with specific rooms and time periods.

3.2. Perceptual Data

For the multisensory data gathering process, each participant session utilized a synchronized multi-modal capture system [11]. The primary capture device was a Meta Quest 3 HMD providing egocentric RGB video at 30fps and 6-DOF head pose via onboard SLAM. Gaze data was captured using integrated eye-tracking hardware, providing 2D gaze position enabling tracking of focus on specific objects [46].

3.3. Study Protocol

Eighteen participants (ages 50–95) from two closely related families participated in the semantic/sensory data collection process. All participants had extensive prior experience with the home, having attended family gatherings there regularly over 30–50 years. Critically, none had visited the home in several years.

Participants were first asked to describe a memory associated with a specific room. They then donned the HMD and navigated the reconstruction while describing memories [5, 42], with original home movies playing on spatially anchored virtual surfaces and the physical memory boxes available for tangible interaction in mixed reality [20]. Session durations ranged from 12 to 28 minutes. All sessions were transcribed and annotated, enabling the linking of verbal memory content to physical and tangible interactions in specific rooms and objects in the reconstruction.

4. Dataset Analysis and Tentative Findings

4.1. Memory Modulates Spatial Attention

When participants navigated the simulation environment, gaze was significantly concentrated on regions corresponding to verbally recalled memory content [34, 46]. For example, participants who described Christmas gatherings exhibited denser fixation around the dining room table and tree placement areas; participants recalling basement activities directed gaze toward stairway affordances and doorways leading to lower floors [42]. An agent trained only on observation data would fail to replicate these anticipatory attention patterns [15], which are critical for natural interaction in human environments.

4.2. Reaching Toward Remembered Objects

Touch interaction data (Seen in Figure 11) with the tangible memory boxes demonstrated that participants exhibited gestures oriented toward virtual object locations associated with their narrated memories [5, 36], even when those objects were only present in the virtual layer with no physical correlate. This finding aligns with prior work on tactile grounding and cross-modal prediction [32, 44].

Component	Scale	Format
3D scanned objects	800+	.obj / .fbx + textures
Rooms reconstructed	18 across 4 floors	Gaussian splat + mesh
Analog photos scanned	5,886	TIFF 600 dpi
Archival film digitized	40+ hours	MP4 24 fps
Total dataset size	∼4 TB	Mixed
Participants	18	Ages 50–95
Sensor modalities	6	Video, pose, gaze, audio, haptic, olfactory

Table 1. Dataset statistics.

Modality	Device	Rate	Signal
Egocentric video	Meta Quest 3	30 fps	RGB + depth
Head pose (6-DOF)	SLAM / IMU	72 Hz	Position + orientation
Gaze	Integrated eye tracking	90 Hz	2D gaze + pupil
Spatial audio	8-mic array	44.1 kHz	Spatial field + VAD
Tangible interaction	Pressure + proximity	100 Hz	Contact + grip
Verbal output	Directional mic	44.1 kHz	Transcribed narration

Table 2. Sensing modalities.

4.3. Cross-Modal Olfactory–Spatial Association

The integration of olfactory stimuli in the form of original fragrances and scented materials such as carpet from the home, contained in physical memory boxes is an integral component of gathering and preserving the multisensory data [31]. A goal of this project is to pair the multisensory data gathered from the physical objects with the virtual objects.

Qualitative data demonstrated that participants who physically handled olfactory artifacts exhibited longer durations of focus in the virtual room spatially associated with those materials [34], compared to participants who did not engage with the tangible elements. This cross-modal effect demonstrates the value of multi-modal sensing systems beyond the visual-proprioceptive baseline common in current datasets [11, 15].

4.4. Memory Quality Under Different Modality Conditions

Verbal transcriptions were analyzed for memory specificity (named individuals, specific objects, references to moments in time) and emotional valence [34, 42], attributes which were impacted by the multisensory data. We believe that sensory-enhanced multi-modal reconstruction environments can serve as effective memory retrieval interfaces, with implications for XR applications and embodied agent design. This research is still in development.

5. Implications for Embodied Intelligence Research

5.1. Training Data for Memory-Aware Spatial Models

Current spatial representation learning assumes agents interact with environments without prior experience [15, 37]. Our dataset enables training and evaluation of models that incorporate episodic memory [5, 36] as a conditioning signal for spatially oriented behavior. The paired control/experimental structure, featuring the same participants, same spatial environment, with and without memory cues, creates a contrastive dataset for learning the differential effect of memory on navigation and focus [42, 46]. Models trained on this data could exhibit anticipatory behavior and exhibit contextual memory.

5.2. Human-to-Robot Skill Transfer

The movement behaviors documented in our dataset, showing humans physically processing object-memory relationships in a mixed virtual-physical space, provides a novel dataset for human-to-robot skill transfer research [7, 16]. In domestic manipulation settings, robots operating in human homes must account for human spatial memory priors [5, 36] that shape expectations of object placement. Our dataset provides ground-truth behavioral data for learning these priors, enabling robot systems to predict where a human expects to find an object based on contextual and historical signals rather than relying solely on current perception [37]. Importantly, our dataset also connects machine vision with robot perception informed by human value and

Figure 9. 8mm and 16mm films, home movies converted into gaussian splats, and hours of recorded video digitized

meaning. For example, fragile objects that have nostalgic value would be handled differently than newer objects [34].

5.3. Cross-Modal Representation Learning

The co-registered multi-modal streams in JDSD, comprised of egocentric video, gaze, pose, eye movement, audio, and tangible interactions [11, 15], provides a unique resource for cross-modal representation learning [31, 44]. The verbal memory transcriptions serve as semantic anchors linking spatial behavior to high-level conceptual content [34, 42], enabling training of models that can ideally learn to connect representations across sensory modalities and language. We believe this is directly relevant to the workshop's focus on cross-modal spatial-temporal alignment.

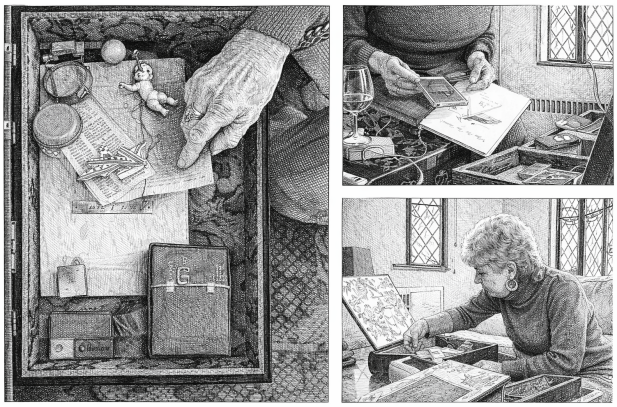

Figure 10. Rendered images from dataset capture showing the sensory integration of touch, sight, and memory during the data gathering process

6. Ethical Considerations

The dataset involves personal data from real individuals interacting with highly personal autobiographical content. Informed consent was obtained from all participants prior to data collection, with explicit opt-in for each data modality through university IRB. Verbal memory narratives were transcribed and reviewed by participants before inclusion. All identifiable information will be pseudonymized in the public release, with faces in egocentric video automatically blurred. The spatial reconstruction of the private home will be released only in anonymized form.

The study design was reviewed to minimize risk of triggering distressing memories. Pre-screening excluded participants with acute grief or trauma history related to the home. A research assistant was present throughout all sessions to monitor participant wellbeing. Two participants became visibly emotional during the VR condition; in both cases the participant voluntarily continued after being offered the opportunity to stop, and post-session interviews indicated positive overall experience.

Datasets of this type raise important questions about privacy in spatial AI training data. Domestic environments are deeply personal, and photorealistic reconstructions at this fidelity create persistent spatial records with potentially sensitive implications. We advocate for community norms around consent, data minimization, and participant control over spatial personal data, and commit to an ethical dataset release.

7. Conclusion

We have presented the Jefferson Drive Spatial Memory Dataset, a large-scale multi-modal corpus capturing human embodied interaction with a photorealistic reconstruction [25, 35] of a personally familiar domestic environment. The dataset uniquely documents how memory [5, 36] modulates spatial attention [46], physical behavior [42], and cross-modal sensory integration [31, 44] in real human participants navigating reconstructed space. Our analysis reveals systematic memory-driven differences in gaze, pose, and verbal output [34, 46] that provide novel training signals for embodied AI systems [15, 37].

The dataset addresses a fundamental gap in current spatial AI benchmarks: the absence of memory-conditioned human-environment data [11, 15, 16]. By capturing how humans navigate space and interact with multisensory stimulation in the context of autobiographical memory [5, 34, 36], we provide data that captures the upstream perceptual and memorial processes that robotic systems must ultimately model [7, 37]. We hope this dataset will stimulate new research in cross-modal spatial understanding, memory-aware embodied agents [5, 42], and multisensory intelligence [31, 44].

References

[1] Way Kiat Bong, Florian Maußer, Margot van Eck, Diogo De Araujo, Jorg Tibosch, Tobias Glaum, and Weiqin Chen. *Designing Nostalgic Tangible User Interface Application for Elderly People*, pages 471–479. Springer International Publishing, 2020. 4

[2] Elise Bonnail, Wen-Jie Tseng, Mark Mcgill, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. Memory manipulations in extended reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023. 4

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging, 2019. 2

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, pages 361–378. Springer, 2020. 2

[5] Paul Byrne, Sean Becker, and Neil Burgess. Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114(2):340–375, 2007. 1, 2, 6, 7, 8

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 2

[7] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots, 2024. 7, 8

[8] Mihaly Csikszentmihalyi and Eugene Halton. *The Meaning of Things: Domestic Symbols and the Self.* Cambridge University Press, 1981. 4

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2

[10] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, pages 1–10, 2018. 2

[11] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *NeurIPS*, pages 13800–13813, 2022. 6, 7, 8

[12] Èric Pairet, Paola Ardón, Xingkun Liu, José Lopes, Helen Hastie, and Katrin S. Lohan. A digital twin for human-robot interaction. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 372–372, 2019. 4

[13] Judith Amores Fernandez, Nirmita Mehra, Bjoern Rasch, and Pattie Maes. Olfactory wearables for mobile targeted memory reactivation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023. 3

[14] Jordan Frith and Jason Kalin. Here, I used to be: Mobile media and practices of place-based digital memory. *Space and Culture*, 19(1):43–55, 2015. 3

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 2, 6, 7, 8

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. EgoExo4D: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, pages 19383–19400, 2024. 2, 7, 8

[17] Michael Grieves. Origins of the digital twin concept, 2016. Unpublished. 4

[18] Morgan Harvey, Marc Langheinrich, and Geoff Ward. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing*, 27:14–26, 2016. 2

[19] Stephan Huber, Renate Berner, Martina Uhlig, Peter Klein, and Jörn Hurtienne. Tangible objects for reminiscing in dementia care. In *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 15–24, 2019. 4

[20] Hiroshi Ishii and Brygg Ullmer. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pages 234–241, 1997. 4, 6

[21] Wenzel Jakob, Marco Tarini, Daniele Panozzo, and Olga Sorkine-Hornung. Instant field-aligned meshes. *ACM TOG*, 34(6), 2015. 5

[22] Jade Jiang, Michael Tobia, Robert Lawther, Dominic Branchaud, and Tomasz Bednarz. Double vision: Digital twin applications within extended reality. In *ACM SIGGRAPH Appy Hour*, 2020. 4

[23] Shiqi Jiang, Zhenjiang Li, Pengfei Zhou, and Mo Li. Memento: An emotion-driven lifelogging system with wearables. *ACM Transactions on Sensor Networks*, 15(1):1–23, 2019. 2

[24] David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020. 4

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 1, 2, 4, 5, 6, 8

[26] Volker Kuchelmeister and Jill Bennett. The amnesia atlas VR: A photographic media interface as memory-prosthesis. In *IEEE Virtual Reality (VR)*, pages 330–330, 2016. 4

[27] Michael Lamming, Peter Brown, Kathleen Carter, Margery Eldridge, Mike Flynn, Gifford Louie, Peter Robinson, and Abigail Sellen. The design of a human memory prosthesis. *The Computer Journal*, 37(3):153–163, 1994. 2

[28] Calvin Lin, Ting-Yu Dai, Aysegul Demir Dilsiz, Drury Crawley, Dev Niyogi, and Zoltan Nagy. UTwin: A digital twin of the UT Austin campus. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 282–283, 2023. 4

[29] David Lindlbauer and Andy D. Wilson. Remixed reality: Manipulating space and time in augmented reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018. 4

[30] Yimeng Liu, Jacob Ritchie, Sven Kratz, Misha Sra, Brian A. Smith, Andrés Monroy-Hernández, and Rajan Vaish. Memento player: Shared multi-perspective playback of volumetrically-captured moments in augmented reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2023. 4

[31] Yiyue Luo, Yunzhu Li, Pratyusha Sharma, Wan Shou, Kui Wu, Michael Foshey, Beichen Li, Tomas Palacios, António Torralba, and Wojciech Matusik. Learning human–environment interactions using conformal tactile textiles. *Nature Electronics*, 4(3):193–201, 2021. 2, 4, 6, 7, 8

[32] Yiyue Luo, Murphy Wonsick, Jessica Hodgins, and Brian Okorn. Tactile embeddings for multi-task learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3348–3355, 2024. 2, 4, 6

[33] Assunta Matassa. Interaction with a personalised smart space to enhance people everyday life. In *DC@CHItaly*, 2015. 3

[34] James L. McGaugh. Memory — a century of consolidation. *Science*, 287(5451):248–251, 2000. 2, 3, 6, 7, 8

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4, 8

[36] John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978. 1, 2, 6, 7, 8

[37] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *NeurIPS*, 2021. 2, 4, 7, 8

[38] David C. Rubin and Sharda Umanath. Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review*, 122(1):1–23, 2015. 2

[39] Corina Sas, Tomasz Fratczak, Matthew Rees, Hans Gellersen, Vaiva Kalnikaite, Alina Coman, and Kristina Höök. AffectCam: Arousal-augmented SenseCam for richer recall of episodic memories. In *CHI Extended Abstracts on Human Factors in Computing Systems*, 2013. 3

[40] Oliver Schreer, Markus Worchel, Rodrigo Diaz, Sylvain Renault, Wieland Morgenstern, Ingo Feldmann, Marcus Zepp, Anna Hilsmann, and Peter Eisert. Preserving memories of contemporary witnesses using volumetric video. *i-com*, 21 (1):71–82, 2022. 4

[41] Yulia Sion, Claudia Diaz Reyes, David Lamas, and Mohammad Mokhalled. Vibmory — mapping episodic memories to vibrotactile patterns. In *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2023. 4

[42] Megan E. Speer, Jamil P. Bhanji, and Mauricio R. Delgado. Savoring the past: Positive memories evoke value representations in the striatum. *Neuron*, 84(4):847–856, 2014. 2, 6, 7, 8

[43] Thomas Suddendorf, Donna Rose Addis, and Michael C. Corballis. Mental time travel and the shaping of the human mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1317–1324, 2009. 2

[44] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019. 2, 4, 6, 8

[45] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020. 2

[46] Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):1–23, 2011. 2, 3, 6, 7, 8

[47] Paolo Testolina, Michele Polese, Pedram Johari, and Tommaso Melodia. BostonTwin: the Boston digital twin for ray-tracing in 6G networks. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 441–447, 2024. 4