Multimodal Perception Methods for Embodied Robots: Sensors, Fusion Architectures, and Human–Robot Interaction Tasks

Leyou Su[1, a], Mingren Zheng[2, b], Zihan Zhao[3, c]

[1] College of the Holycross, 1 College Street, Worcester, MA 01610, USA

[2] Tabor Academy, 66 Spring Street, Marion, MA 0273, USA

[3] Kimball Union Academy, 7 Campus Center Drive, Meriden NH, 03770, USA

[a]13718661061@163.com, [b]Mingrenzheng2007@gmail.com, [c]little.brand.6@outlook.com

Abstract

Embodied robots must perceive geometry, motion, contact, human behavior, and task context under latency and safety constraints that no single sensor can satisfy. Multimodal perception addresses this limitation by combining complementary observations, but adding modalities also introduces calibration, synchronization, representation, missing-data, and confidence-management problems. This paper presents a method-centered review organized around the chain from sensing to representation and alignment, fusion architecture, interaction task, and failure handling. We examine frame and event vision, depth and ranging, audio, human-state cues, force/torque sensing, and tactile sensing; then compare data-level, feature-level, decision-level, and temporal fusion. The review connects these choices to scene and object understanding, affordance and physical-property estimation, human activity recognition, and intent or task-stage prediction. A central finding is that fusion depth should follow the strength and stability of cross-modal coupling. Early fusion can preserve fine correspondence when sensors are well calibrated; intermediate fusion supports learned cross-modal interactions; late fusion offers modularity and graceful degradation; and temporal state estimation is essential when observations arrive asynchronously. Robust systems should represent sensor validity and uncertainty explicitly, train with realistic modality dropout and corruption, and separate semantic inference from safety-critical constraints. Foundation models improve open-vocabulary grounding and task communication, but do not eliminate timing, observability, or verification requirements. The paper concludes with a condition–method–benefit–cost guide and identifies benchmark practices needed for deployment-relevant evaluation: synchronized raw data, missing-modality protocols, calibration perturbations, end-to-end latency, uncertainty calibration, and task-level consequences of perceptual error.

Keywords

Multimodal perception, sensor fusion, embodied robotics, human–robot interaction, tactile sensing, intent prediction.

1. Introduction

Physical interaction makes perception consequential. A service robot that confuses a reflective surface with free space may collide; a manipulator that recognizes an object but misjudges friction may drop it; a collaborative robot that detects a hand but misunderstands the worker's task stage may interrupt or endanger the workflow. Embodied intelligence therefore requires more than semantic recognition. It requires sufficiently accurate, timely, and uncertainty-aware estimates of a world that changes partly because the robot acts.

No individual sensing modality provides all needed information. Cameras offer rich appearance and spatial cues but are sensitive to illumination, occlusion, and viewpoint. Depth cameras and LiDAR support geometry but face range, surface, and sparsity limitations. Microphones capture speech and acoustic events while also capturing reverberation and noise. Force, torque, and tactile sensors reveal contact that external vision may not observe, but only after interaction and over limited areas. Human pose, gaze, facial behavior, and motion provide evidence about attention and intent, yet these signals are ambiguous and socially variable. Multimodal perception seeks to exploit complementarity while controlling the new failure modes created by fusion.

The multimodal machine-learning literature organizes core problems around representation, alignment, fusion, translation, and co-learning [1]. Robotics adds hard real-time constraints, changing viewpoints, active sensing, embodiment-specific geometry, physical contact, and safety. Surveys of human–robot multimodal perception document a broad range of sensors and interaction tasks [2], [3], while human–robot collaboration research emphasizes the coupling among perception, action, and work design [4]. A recent synthesis further connects multimodal perception to downstream HRI decision making [15]. Foundation-model robotics extends open-vocabulary grounding and the interface between language, vision, and action [5]–[7], [16], but does not remove the need for calibration, state estimation, and failure management.

This review addresses four questions. First, what complementary information and characteristic failures do major sensing modalities provide? Second, how do early, intermediate, late, and temporal fusion architectures trade information retention, synchronization burden, modularity, and robustness? Third, how do those choices map to scene understanding, affordance estimation, human activity recognition, and intent prediction? Fourth, which design and evaluation practices improve robustness to missing modalities, distribution shift, latency, and uncertainty?

The paper contributes an integrated sensing-to-task taxonomy and a practical selection guide. The organizing principle is that fusion should be designed around the causal and temporal structure of the task rather than around sensor availability alone. The intended result is a review that helps system designers reason about when modalities should interact, what is lost at each fusion stage, and how perceptual uncertainty reaches robot behavior.

2. Review Scope and Perception Pipeline

The review is a structured narrative synthesis of primary robotics papers, established multimodal and human–robot-interaction surveys, and representative foundation-model reports. It emphasizes embodied perception for manipulation and human–robot collaboration. Autonomous driving and purely virtual multimodal learning are included only where a method transfers directly to robotic sensing. The sources span classical sensor fusion, learned representation, tactile sensing, event vision, human-intent inference, and missing-modality robustness.

We model the perception pipeline as five coupled layers:

This pipeline exposes two often-hidden distinctions. First, preprocessing and alignment are inference steps, not neutral plumbing. A depth completion algorithm or pose tracker can introduce structured errors that fusion later treats as measurements. Second, the target representation should be defined by action. A visually rich scene description is not sufficient if the controller needs contact state within milliseconds; a precise force trace is not sufficient if the task planner needs the human partner's goal.

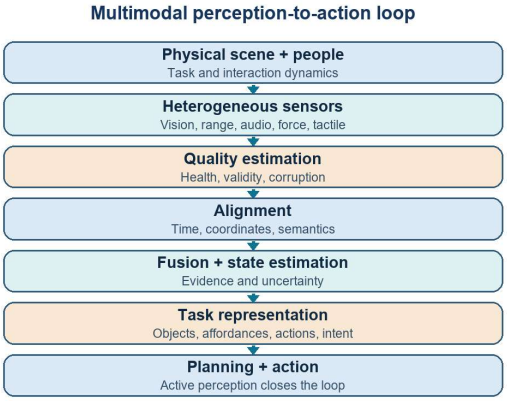


Figure 1. Multimodal embodied-perception pipeline: physical scene and human activity → heterogeneous sensors → preprocessing and quality estimation → temporal/spatial/semantic alignment → data-, feature-, decision-, or state-level fusion → task representation → planning and action. Robot motion changes both the scene and future observability, while runtime monitors return validity and uncertainty signals to fusion and control.

3. Sensing Modalities and Complementarity

3.1. Frame Vision, Depth, and Event Cameras

RGB cameras remain central because they capture texture, color, object identity, human appearance, and broad scene context at low hardware cost. Stereo or RGB-D systems add geometry, supporting localization, obstacle avoidance, grasp planning, and body tracking. Their weaknesses are partly independent: passive stereo degrades on low-texture surfaces, structured-light and time-of-flight sensors can suffer from range or interference limits, and all can be affected by occlusion and viewpoint. Sensor placement therefore determines what the system can observe, especially during manipulation when the arm or object blocks an external camera.

Event cameras report changes in log intensity asynchronously rather than complete frames. Their high temporal resolution, low latency, and high dynamic range make them attractive for fast motion and difficult lighting, while their sparse output and different noise model require specialized processing [8]. Event vision complements rather than replaces frame vision: frames supply appearance and absolute intensity context, while events supply precise change information. Fusion is valuable only if timestamps and motion compensation are sufficiently accurate.

3.2. LiDAR, Radar, and Proprioception

LiDAR provides metric range and supports mapping, navigation, and geometric scene representations. Compared with image features, point clouds are sparse and may have limited semantic content, but they can remain useful across illumination changes. Radar measures range and relative velocity and can operate through some visual degradation, although angular resolution and multipath complicate interpretation. In mobile systems, these exteroceptive sensors are fused with wheel odometry and inertial measurements because geometry alone may not reveal short-term motion reliably.

Proprioceptive signals-joint positions, velocities, motor currents, inertial measurements, and actuator temperatures-describe the robot's own state. They are essential for transforming external observations into body coordinates and for detecting disagreement between intended and achieved motion. Although sometimes excluded from “multimodal perception,” proprioception closes the embodiment loop. A camera may indicate that an object moved; joint

and force signals help determine whether the robot caused the movement, lost contact, or was obstructed.

3.3. Audio and Human-state Cues

Microphone arrays support speech, speaker direction, acoustic event detection, and sometimes material or contact inference. Audio can remain informative when the relevant event is outside the camera view, but reverberation, machinery, overlapping speech, and privacy concerns are central. Speech recognition output is a derived, uncertain modality. Treating it as error-free text hides timing, accent, and noise failures.

Human pose, hand configuration, head orientation, gaze, facial behavior, and trajectories can indicate action and attention. These cues are usually estimated from vision and depth rather than sensed independently, so correlated failure matters. A face detector and gaze estimator built on the same occluded image do not constitute two independent confirmations. Physiological or wearable signals can add information but increase intrusiveness and deployment complexity. Multimodal HRI surveys emphasize that social and task context affects interpretation [2], [4].

3.4. Force, Torque, and Tactile Sensing

Wrist force/torque sensors estimate interaction wrench and support compliant insertion, contact detection, handover, and collision response. Joint torque or motor-current estimates extend coverage but include friction and model uncertainty. These signals are fast and directly connected to physical interaction, yet they reveal limited spatial detail about where contact occurs.

Tactile arrays and vision-based tactile sensors provide local pressure, deformation, slip, geometry, or shear cues. GelSight demonstrated high-resolution surface geometry and force-related sensing through an elastomer observed by a camera [9]. DIGIT reduced the size and cost of this class of sensor, facilitating in-hand manipulation research [10]. Recent work combining tactile and visual information in dexterous housekeeping tasks illustrates the value of contact-rich fusion for deformable and occluded objects [11]. Tactile sensing is nevertheless local, subject to wear, and available only after contact. Calibration may change with elastomer aging, temperature, and mounting.

Table 1. Complementary information and characteristic failures

Modality	Strongest information	Characteristic weaknesses	Useful complement
RGB/frame vision	Appearance, semantics, human cues	Illumination, occlusion, viewpoint	Depth, events, proprioception
RGB-D/LiDAR	Metric geometry and free space	Range limits, sparsity, reflective/absorptive surfaces	RGB, inertial sensing
Event vision	Fast change and high dynamic range	Sparse output, specialized noise and processing	Frames, inertial sensing
Audio	Speech, direction, off-screen events	Noise, reverberation, privacy	Vision, task state
Proprioception	Body configuration and motion	Drift, model dependence, limited environment information	External geometry, force
Force/torque	Interaction wrench and contact onset	Weak contact localization, structural coupling	Vision, tactile arrays
Tactile	Local geometry, slip, pressure, shear	Contact-only, wear, limited coverage	Vision, proprioception

4. Representation, Alignment, and Fusion

4.1. Quality Estimation and Preprocessing

Fusion begins before features meet. Raw measurements should carry timestamps, calibration identifiers, exposure or gain state, and health information. Preprocessing may include denoising, distortion correction, point-cloud filtering, speech enhancement, event accumulation, tactile normalization, or coordinate transformation. The output should include a validity estimate where possible. Saturated tactile pixels, dropped frames, and stale transforms are not ordinary observations and should not be embedded as if they were.

Quality can be estimated from sensor diagnostics, signal statistics, model residuals, or cross-modal consistency. These sources answer different questions. A camera can be internally healthy while its view is occluded; a depth map can appear dense while being geometrically biased; modalities can agree because they share the same upstream error. A robust design keeps hardware health, observation quality, and semantic confidence distinct.

4.2. Temporal Alignment

Sensors sample at different rates, with different transport delays and processing time. A force signal may arrive at kilohertz rates, a camera at tens of frames per second, and a language interpretation after hundreds of milliseconds. Timestamping at capture is preferable to timestamping after processing. Fusion can use resampling, buffering, interpolation, asynchronous filters, or event-driven updates. The appropriate method depends on task dynamics: handover contact may require millisecond-scale force updates, whereas verbal intent can update a slower task belief.

Temporal misalignment creates physically plausible but false associations. A hand position from one instant fused with force from another can reverse the inferred order of contact. Fixed offsets are insufficient when compute and network delay vary. Systems should measure end-to-end latency distributions and either predict observations to a common time or maintain a state estimator that accepts asynchronous updates.

4.3. Spatial and Semantic Alignment

Spatial alignment maps observations into common coordinates using intrinsic and extrinsic calibration, kinematics, and localization. Errors grow when cameras move, robot links flex, or tactile sensors are remounted. Online calibration or residual monitoring is useful where geometry changes. Semantic alignment maps representations with different meanings-for example, spoken “that cup,” a gaze ray, an image region, and a candidate object in a task graph. This problem cannot be solved by geometric calibration alone.

Cross-attention and shared latent spaces can learn semantic correspondence, but they depend on representative paired data. The classical multimodal taxonomy warns that alignment may be explicit or implicit and may occur at different temporal scales [1]. In robotics, causal structure helps: robot motion, contact, and task outcomes provide supervision. Active perception can deliberately move a camera or touch an object to resolve ambiguity rather than passively accepting uncertain alignment.

4.4. Data-level Fusion

Data-level or early fusion combines raw or minimally processed measurements. Examples include stacking synchronized images and depth maps, projecting point clouds into an image, or combining taxels and force readings in a contact patch. Early fusion retains fine-grained correspondence and allows a model to learn cross-modal patterns from the start. Its costs are strict calibration, high bandwidth, sensitivity to missing channels, and architecture coupling.

Early fusion is appropriate when modalities are tightly synchronized, spatially registered, and jointly needed at local resolution. RGB-D segmentation is a typical case. It is less attractive when a slow semantic signal is paired with a fast control signal or when sensors are optional across product variants. Training-time corruption and explicit masks can reduce brittleness, but cannot reconstruct information that was never observed.

4.5. Feature-level Fusion

Intermediate fusion encodes each modality and then exchanges or combines features. Concatenation, gated mixtures, tensor fusion, graph networks, and cross-attention are common mechanisms. Separate encoders respect modality-specific statistics; shared layers learn complementary interactions. The designer chooses the layer at which semantic abstraction and spatial detail are balanced.

Feature fusion is powerful for vision–language grounding, audio-visual event recognition, and vision–tactile manipulation. It can also obscure calibration: attention weights are not necessarily reliable explanations, and a model may exploit dataset shortcuts. Feature dimensionality, token count, and memory affect control latency. Encoders should preserve timestamps and coordinate references when the downstream task needs geometry rather than only classification.

4.6. Decision-level Fusion

Late fusion combines predictions or beliefs from modality-specific models. Voting, weighted averaging, Bayesian combination, Dempster–Shafer variants, or learned gating can be used. This design is modular and accommodates different sampling rates, training sets, or product configurations. It also supports graceful degradation: if audio fails, a vision-only intent estimate can continue with lower confidence.

The limitation is information loss. Independent classifiers may discard subthreshold evidence that would become meaningful jointly. Probability combination is valid only when calibration and dependence assumptions are addressed. Two confident models trained on correlated labels and scenes should not be treated as independent evidence. Late fusion is strongest when component outputs have explicit semantics and well-characterized uncertainty.

4.7. Temporal and State-level Fusion

Robotic perception often estimates a latent state that persists through time: object pose, contact mode, human goal, or task phase. Filters and sequence models integrate asynchronous evidence with dynamics. Differentiable filters combine learned components with recursive state estimation, enabling end-to-end optimization while retaining a probabilistic structure [12]. Recurrent networks and transformers offer flexible temporal context but require careful latency and memory control.

Table 2. Fusion-stage selection guide

Condition	Preferred starting point	Principal benefit	Main cost or risk
Tight time/space registration and dense correspondence	Early fusion	Preserves local cross-modal detail	Calibration and missing-channel brittleness
Complementary semantics with modality-specific encoders	Intermediate fusion	Learns rich interactions	Data, compute, opacity, shortcut learning
Modular subsystems or optional sensors	Late fusion	Replaceability and graceful degradation	Information loss and confidence dependence
Asynchronous evidence about a persistent state	Temporal/state fusion	Handles dynamics and occlusion	Model mismatch and delayed updates

State-level fusion is especially important under occlusion. When an object disappears behind a hand, the system should propagate a belief rather than reset the scene. The dynamics model must allow discontinuities such as a grasp or handover. For safety-critical decisions, the state estimator should expose uncertainty and innovation residuals to the controller.

5. Perception Tasks in Human–Robot Interaction

5.1. Scene and Object Understanding

Scene understanding estimates objects, people, surfaces, free space, relations, and changes. Geometry and semantics should remain linked: an open-vocabulary label without a reliable pose cannot guide grasping, while an accurate point cluster without task meaning cannot guide planning. Foundation models expand semantic queries and instruction grounding [5]–[7], but deployment requires local object state, access constraints, and update timing. Memory is useful for partially observed environments, provided stale observations are identified.

For collaboration, a scene representation should include ownership, reachability, human work zones, and the consequences of motion. Uncertainty should affect behavior. A robot that is unsure whether a region contains a transparent obstacle should change viewpoint or slow down rather than collapse the estimate to "free."

5.2. Affordance and Physical-Property Estimation

Affordances describe action possibilities relative to an agent, object, and task. Visual appearance can suggest a handle, container, or deformable region; depth supplies geometry; tactile and force signals verify contact, compliance, friction, fill state, or slip. The problem is naturally sequential. Before contact, the robot holds a visual hypothesis; during approach, geometry narrows pose uncertainty; after contact, tactile evidence updates the physical model.

This structure favors intermediate fusion for planning and fast state-level fusion for control. Mao et al. show how multimodal tactile sensing fused with vision can support dexterous housekeeping tasks [11]. Such results demonstrate the value of complementary contact cues, not universal generalization. Evaluation should report object novelty, material variation, sensor wear, and recovery from incorrect affordance hypotheses.

5.3. Human Action and Activity Recognition

Action recognition estimates short behaviors such as reaching or pointing; activity recognition incorporates longer task context. Vision supplies pose and objects, audio supplies speech and events, and robot or environment sensors indicate shared-object state. Temporal order matters: reaching toward a part after requesting it has a different meaning from moving the same hand during an unrelated assembly step.

Human–robot perception should be evaluated across people, clothing, body shapes, viewpoints, cultures, and accessibility conditions. A model trained on staged gestures may fail on natural work. Privacy-preserving representations can reduce retention of identity-rich video, but derived skeletal or embedding data may still be sensitive. The minimum necessary sensing principle is therefore both an ethical and engineering design rule.

5.4. Intent and Task-stage Prediction

Intent is latent and cannot be directly measured. It is inferred from behavior, context, dialogue, and task structure. A recent review of intent inference and action prediction in collaboration emphasizes the range of model-based and learning-based approaches and the importance of temporal prediction [13]. Multimodal systems can combine gaze, body motion, speech, object state, and workflow history. Yet ambiguity is irreducible: people can change their minds, behave strategically, or deviate from the modeled task.

Robots should treat intent as a belief with alternatives, not as a ground-truth label. High-consequence actions require confirmation, conservative motion, or an opportunity for correction. Prediction horizon should match the action: short-horizon collision avoidance differs from inferring a worker's goal for cooperative planning. Evaluation should score not only prediction accuracy but calibration, time-to-correct-prediction, and the cost of false anticipation.

6. Robustness, Uncertainty, and Design Guidance

6.1. Missing and Degraded Modalities

Modalities fail in structured ways: cameras are occluded, microphones are masked by machinery, depth returns disappear on surfaces, networked sensors become stale, and tactile skins degrade. Training with random modality dropout is useful, but independent random masks may not resemble real failures. Robust multimodal prompting research shows the continuing difficulty of inference with missing modalities [14]. Robotics benchmarks should therefore include correlated outages, calibration shifts, latency spikes, and quality degradation.

A practical architecture has unimodal minimum-capability paths, explicit validity masks, and policies tied to operating modes. Loss of speech may permit task completion through a display; loss of depth may require slower motion; loss of safety-rated sensing may require a protective stop. The fallback should be verified rather than improvised by a generative model.

6.2. Uncertainty and Confidence Calibration

Uncertainty has at least three sources: sensor noise, model uncertainty, and ambiguity inherent in the scene or human behavior. Confidence scores from different models are rarely comparable without calibration. Selective prediction allows the system to abstain, request another view, ask a question, or transfer control. Cross-modal disagreement is a useful signal, but agreement does not prove correctness when errors are correlated.

Calibration should be evaluated under distribution shift, not only on held-out samples from the same collection. Metrics such as expected calibration error can summarize confidence behavior, but task-weighted risk is more meaningful for control. A false “clear path” prediction has a different cost from an uncertain object label. The fusion layer should pass risk-relevant uncertainty to planning rather than converting every state into a point estimate.

6.3. Latency and Compute Allocation

Multimodal transformers and foundation models can be computationally expensive. The system should allocate computation according to control frequency. Fast collision, balance, and contact loops should not depend on slow cloud inference. Higher-level semantic models can operate asynchronously, producing goals or constraints consumed by local controllers. Caching and event-triggered inference can reduce load, but stale semantic outputs need timestamps and invalidation rules.

This multi-rate design separates semantic richness from safety reaction without isolating them. A high-level model can request an active-perception action, while low-level monitors bound that action. End-to-end latency must include sensor exposure, transfer, preprocessing, inference, fusion, planning, and actuation-not model inference alone.

6.4. Evaluation Priorities

Current evaluations often report unimodal or clean multimodal accuracy. Deployment-oriented benchmarks should release synchronized raw streams, calibration records, sensor health, and task outcomes. They should define missing-modality and corrupted-modality protocols; perturb time and extrinsic calibration; report compute, energy, and end-to-end latency;

evaluate uncertainty calibration; and connect perception errors to task completion, intervention, and safety margins.

Generalization splits should separate environments, objects, people, sensors, and embodiments. The Open X-Embodiment program shows the potential of heterogeneous data for cross-robot learning [7], while also highlighting that heterogeneous embodiments create representation and control challenges. Dataset scale should therefore be paired with documentation of coverage and failure, not treated as an independent measure of robustness.

Table 3. Failure-to-mitigation mapping

Failure	Observable indicator	Design response	Evaluation requirement
Temporal mismatch	Innovation spikes; inconsistent motion	Capture timestamps; asynchronous state update	Inject variable delays
Calibration drift	Spatial residuals grow	Online checks; recalibration; conservative bounds	Perturb extrinsics/intrinsics
Missing modality	Health flag or stale stream	Masked fusion; verified fallback	Correlated dropout scenarios
Correlated model error	Confident but shared failure	Diverse sensing; causal checks; active perception	Cross-domain and counterfactual tests
Miscalibrated confidence	Error rises without uncertainty	Calibration; abstention; task-weighted thresholds	Reliability curves under shift

7. Conclusion

Multimodal perception enables embodied robots to combine semantics, geometry, dynamics, contact, and human context, but it also creates synchronization, calibration, dependence, and uncertainty problems. Frame and event vision, ranging, audio, proprioception, force, and tactile sensing provide complementary evidence with distinct observation windows and failure modes. Early fusion is strongest for stable dense correspondence, intermediate fusion for learned semantic interaction, late fusion for modularity, and temporal fusion for asynchronous evidence about persistent state. The correct choice follows task dynamics and failure consequences rather than a universal hierarchy.

For scene understanding, affordance estimation, human activity recognition, and intent prediction, robust perception requires explicit timestamps, sensor validity, calibrated uncertainty, and action-aware evaluation. Foundation models broaden language and semantic grounding, yet safety-critical execution still needs bounded controllers and verified fallback. Future progress should be measured by performance under missing modalities, calibration perturbation, distribution shift, and latency-not only by clean-data accuracy. A multimodal robot is commercially and socially useful when it knows not just what its sensors suggest, but when those suggestions are too late, too uncertain, or too incomplete to act upon.

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[2] Duncan, J. A., Alambeigi, F., & Pryor, M. W. (2024). A survey of multimodal perception methods for human–robot interaction in social environments. ACM Transactions on Human-Robot Interaction, 13(4), Article 46, 1–50. https://doi.org/10.1145/3657030

[3] Mohd, T. K., Nguyen, N., & Javaid, A. Y. (2022). Multi-modal data fusion in enhancing human-machine interaction for robotic applications: A survey [Preprint]. arXiv. https://arxiv.org/abs/2202.07732

[4] Bauer, A., Wollherr, D., & Buss, M. (2008). Human–robot collaboration: A survey. International Journal of Humanoid Robotics, 5(1), 47–66. https://doi.org/10.1142/S0219843608001303

[5] Driess, D., Xia, F., Sajjadi, M. S. M., et al. (2023). PaLM-E: An embodied multimodal language model [Preprint]. arXiv. https://arxiv.org/abs/2303.03378

[6] Brohan, A., Brown, N., Carbajal, J., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control [Preprint]. arXiv. https://arxiv.org/abs/2307.15818

[7] Open X-Embodiment Collaboration. (2023). Open X-Embodiment: Robotic learning datasets and RT-X models [Preprint]. arXiv. https://arxiv.org/abs/2310.08864

[8] Gallego, G., Delbrück, T., Orchard, G., et al. (2022). Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1), 154–180. https://doi.org/10.1109/TPAMI.2020.3008413

[9] Yuan, W., Dong, S., & Adelson, E. H. (2017). GelSight: High-resolution robot tactile sensors for estimating geometry and force. Sensors, 17(12), Article 2762. https://doi.org/10.3390/s17122762

[10] Lambeta, M., Chou, P.-W., Tian, S., et al. (2020). DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. IEEE Robotics and Automation Letters, 5(3), 3838–3845. https://doi.org/10.1109/LRA.2020.2977257

[11] Mao, Q., Liao, Z., Yuan, J., & Zhu, R. (2024). Multimodal tactile sensing fused with vision for dexterous robotic housekeeping. Nature Communications, 15, Article 6871. https://doi.org/10.1038/s41467-024-51261-5

[12] Lee, M. A., Yi, B., Martín-Martín, R., Savarese, S., & Bohg, J. (2020). Multimodal sensor fusion with differentiable filters. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 10444–10451). https://doi.org/10.1109/IROS45743.2020.9341579

[13] Hoffman, G., Bhattacharjee, T., & Nikolaidis, S. (2024). Inferring human intent and predicting human action in human–robot collaboration. Annual Review of Control, Robotics, and Autonomous Systems, 7, 73–95. https://doi.org/10.1146/annurev-control-071223-105834

[14] Jang, J., Wang, Y., & Kim, C. (2024). Towards robust multimodal prompting with missing modalities. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8070–8074). https://doi.org/10.1109/ICASSP48485.2024.10447257

[15] Zhao, W., Gangaraju, K., & Yuan, F. (2025). Multimodal perception-driven decision-making for human–robot interaction: A survey. Frontiers in Robotics and AI, 12. https://doi.org/10.3389/frobt.2025.1604472

[16] Xu, Z., Wu, K., Wen, J., Li, J., Liu, N., Che, Z., & Tang, J. (2024). A survey on robotics with foundation models: Toward embodied AI [Preprint]. arXiv. https://arxiv.org/abs/2402.02385