Generating Human-Understandable Descriptions of Novel Objects for Verbal Interactions with Edge-Based Robots

Sarah Schneider[1,2], Evan Krause[1], Marlow Fawn[1], Doris Antensteiner[2], Csaba Beleznai[2], Daniel Soukup[2], and Matthias Scheutz[1]

[1] Department of Computer Science, Tufts University, Medford, MA 02155, USA
[2] Center for Vision, Automation and Control, AIT Austrian Institute of Technology, Vienna 1210, Austria
{sarah.schneider, evan.krause, marlow.fawn, matthias.scheutz}@tufts.edu
{doris.antensteiner, csaba.beleznai, daniel.soukup}@ait.ac.at

Abstract. Mobile robots are becoming increasingly prevalent across a wide range of environments. They must effectively perceive the open world despite constraints in computational power and network resources, while also communicating their understanding to human partners. We present a compact neural structural encoder that supports object-level open-world understanding by decomposing novel objects into a set of known primitives drawn from a component vocabulary. Embedded within a cognitive architecture, the system maps geometric information into human-language descriptions and visualizations that prioritize structured interpretability over unrestricted expressiveness. Our approach uses synthetic data generation, model training on synthetic data, and reconstruction consistency estimation to indicate description reliability. A user study confirms that the generated descriptions are informative for human collaborators and shows how our human-language descriptions compare to GPT-generated descriptions, which rely on far greater computational resources. Different description versions are compared based on user preferences, and an on-robot demonstration illustrates the practical feasibility of our method. This work serves as a blueprint for an efficient and accessible vision-based object description system suited for open-world robotic collaboration.

Keywords: Representation Learning · Computer Vision for HRI · Interpretable AI

1 Introduction

Vision-Language Models (VLMs) show exceptional capabilities in tasks that involve describing visual scenes [31, 17, 19, 55]. Despite these advancements, VLMs still face limitations in reasoning about spatial arrangements of object parts and their mereological relationships, including how parts are structured, composed, and contained within the overall object [6]. Many of these models are hosted

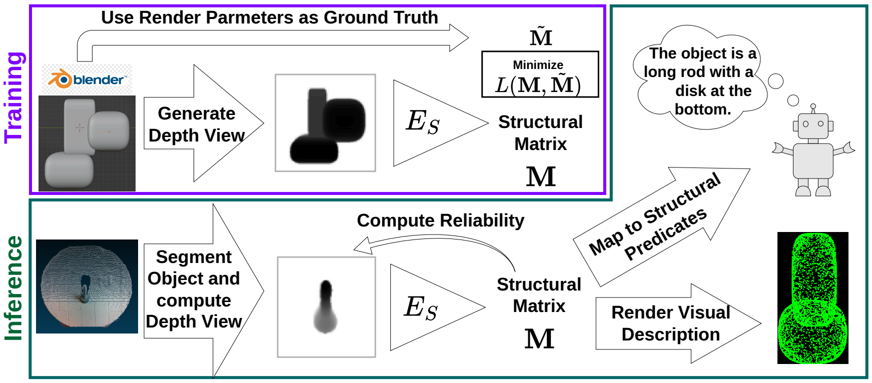

Training
Use Render Parmeters as Ground Truth
$\tilde{\mathbf{M}}$
blender
Generate Depth View
E_S
Minimize
$L(\mathbf{M}, \tilde{\mathbf{M}})$
Structural Matrix
$\mathbf{M}$
The object is a long rod with a disk at the bottom.
Inference
Segment Object and compute Depth View
Compute Reliability
E_S
Structural Matrix
$\mathbf{M}$
Map to Structural Predicates
Render Visual Description

Fig. 1: We train a structural encoder E_s on synthetic objects to obtain a compact representation $\mathbf{M}$, which we map to human language and visual descriptions of real-world objects.

on remote servers, requiring reliable network access. When deployed locally on robots, they demand substantial computational resources [17, 56], which can be undesirable due to the high power demands placed on the robot's battery capacity [10].

Autonomous robots that have to operate in task environments for extended periods of time without intermittent recharging cannot afford to have their batteries drained quickly because of power-hungry foundation models. When robots communicate with remote human operators about task-related objects without live video feeds (due to low throughput, etc.), natural language descriptions become valuable, as they convey one's understanding of the environment [52]. Trust directly affects people's willingness to accept robot-generated information [28], and a robot's ability to transparently describe novel objects, i.e., objects it has not encountered before, is critical for successful human-robot collaboration. Moreover, traditional AI systems often assume a closed world in which all relevant concepts are known beforehand, and the system model is considered complete, leaving robots ill-prepared to meaningfully characterize or communicate about unfamiliar objects encountered in open-ended environments [22, 9, 30]. The research challenge thus is two-fold: (1) to develop a vision processing model that produces compact human-interpretable representations of novel objects, and (2) expressing these representations in natural language so that the robot can effectively communicate its perceptions.

We tackle both challenges with a compositional geometry-based representation pipeline. Specifically, a lightweight convolutional neural network (CNN) maps depth observations of an object to an interpretable structural representation that parameterizes the object as a set of superellipsoidal primitives, a compact mathematical representation capable of modeling a wide range of shapes [3]. A rule-based natural language module leverages this representation to facilitate

human-robot communication about novel objects. In addition, these representations can be rendered as language-independent visual descriptions, allowing humans with different communication preferences to intuitively understand the object. The reliability of the descriptions can be assessed by comparing the spatial representation of the reconstructed object with its point cloud.
Our primary goal is to establish a robotic system that effectively interacts with a human user. Accordingly, we demonstrated the meaningfulness of our generated descriptions through a user study, where participants successfully matched them to their corresponding objects. To contextualize these results, we included GPT, relying on substantially greater computational resources, as an upper-bound reference. Participants matched our descriptions at a level comparable to GPT, and in some instances, even more successfully. We further compared model and representation size against representative image encoders and compression techniques, and showcased the practical feasibility of our approach with an on-robot demonstration. The proposed approach is designed to be accessible, requiring minimal hardware and training efforts, and its modular structure allows for straightforward adaptation without dependence on cloud-based infrastructure.

2 Motivation and Related Work

Our goal of examining compound geometric objects is strongly motivated by the study of *compositionality*, the design rule guiding the composition of increasingly complex objects from re-usable parts. A key appeal of compositionality lies in its ability to support out-of-distribution generalization while enabling efficient and diverse representations [47, 14]. The decomposition of complex objects into simple parts has been a research endeavor from early on, as findings from neuroscience also indicated an organizational hierarchy at multiple levels [38]. Statistical modelling [57, 40, 46] attempts to capture part hierarchies from images. Deep distributed representations exhibit a large representational capacity and parameter-efficient encodings, but compositionality is not explicitly set as an objective. Some approaches have been proposed [48, 45] to remedy this deficiency, but this research direction still represents an open issue.

Robots that clarify their actions can enhance human-agent collaboration by increasing trust and safety [2]. Concept bottleneck models [33] use interpretable intermediate concepts for further predictions. Others optimize for completeness and interpretability [54], or learn concepts using gradient and global search [44]. Several studies decompose shapes into geometric primitives. Reinforcement and imitation learning have been used to model shapes with primitive-based representations [36]. Superquadrics and Gaussians are used to learn part-aware scenes from multi-view inputs [20, 32]. Optimization-based methods reconstruct objects from multiple views [1], while superquadrics are extracted from geometric features in affordance point clouds [37]. SuperDec [16] leverages superquadrics to produce expressive 3D scene representations by decomposing individual objects.

The motivation for our work extends into disaster response scenarios, where robots help to monitor hazardous environments. Whether in natural (e.g., earthquakes, floods) or man-made crises (e.g., industrial or nuclear emergencies), robots can reduce costs while allowing human operators to remain at a safe distance [7, 8]. These scenarios often involve unfamiliar objects and challenging network conditions. Signal transmission may be hindered by collapsed walls and electromagnetic interference from damaged infrastructure [49, 12, 8]. In such situations, robots must relay observations to their human collaborators without relying on high data transfer or extensive computational resources, given their limited battery capacity [10]. Similar challenges arise in space operations, where robotic systems navigate unexplored environments with limited prior data and constrained communication bandwidth. The absence of large-scale datasets hinders effective training and validation of deep neural networks prior to deployment. Radiation-hardened processors, crucial for resilience in harsh space conditions, are extremely performance-limited, trailing generations behind commercial processors [23, 42].

We propose an interpretable and efficient method for robotic object description under limited hardware and network conditions (see Fig. 1). We create a synthetic dataset of objects built from geometric primitives and train a model to infer their geometric parametrization from depth images using rendering parameters as ground truth. The objects in our dataset are designed to be as fundamental as possible, resembling "object templates". Our data generation process is easily adjustable and does not require any data labeling, while our training procedure avoids the need for extensive resources. Our structural encoder model is designed to be suitable for on-robot deployment.

We integrate the trained model into a cognitive robotic architecture, translating extracted parameters into human-language descriptions that include object attributes (e.g., shape, eccentricity) and spatial relations (e.g., on top, to the left). Alternatively, we can generate schematic visualizations, offering a visual description of the objects. For individuals who have difficulty understanding human-language, visual representations provide an informative alternative, while verbal descriptions assist individuals with visual impairments. The geometric parameters estimated by the structural encoder model provide a highly compact and interpretable representation of objects. The model's pre-defined parametric structure provides a means to assess the reliability of the generated description by comparing the reconstructed objects with their corresponding point clouds.

Overall, the method requires modest hardware and training resources, does not depend on cloud infrastructure, and can be adapted through its modular design to different object structures, interaction needs, and languages.

3 The Structural Encoder Model

This section introduces the proposed structural encoder model and the synthetic data representation used to train it. We first describe how objects are constructed from superellipsoidal components to generate controlled, interpretable training data. We then detail how these components are encoded into a structural matrix, followed by the training procedure of the encoder and the computation of a reconstruction-consistency–based reliability score of the predicted structural representation.

3.1 Superellipsoids as Object Components for Synthetic Data Generation

We represent objects as composites of superellipsoids, as they provide a flexible yet compact mathematical parametrization capable of capturing a wide range of shapes. Each superellipsoid defines a distinct component of the object within a Cartesian coordinate system. Every object contains at least one superellipsoid, referred to as the base component, and optional side components. For a set of points $(x, y, z) \in \mathbb{R}^3$, the surface of a superellipsoid centered at (x_0, y_0, z_0) is defined by [3]:

$$f(x,y,z) = \left(\left(\frac{x - x_0}{a_x} \right)^{\frac{2}{\epsilon_2}} + \left(\frac{y - y_0}{a_y} \right)^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left(\frac{z - z_0}{a_z} \right)^{\frac{2}{\epsilon_1}} = 1. \qquad (1)$$

The main axes of the superellipsoid, $a_x, a_y, a_z \in \mathbb{R}^+$, determine its size along the x-, y-, and z-axes, respectively. The shape parameters $\epsilon_1, \epsilon_2 \in \mathbb{R}^+$ control the “edginess” along the z-axis and the x-y plane, respectively. Smaller values of ϵ_1 and ϵ_2 produce more box-like shapes, while larger values yield smoother, more rounded surfaces. The objects are generated in Blender [4]. The number of components, values for main axes a_x, a_y, a_z, edginess ϵ_1 and ϵ_2 and center coordinates x_0, y_0, z_0 of the superellipsoids are randomly selected from a uniform distribution. Side components are positioned such that they are adjacent to the base component. We consider the base component to be the component at the most centered position. To introduce position invariance, we shift the objects by a random offset. Depth views are generated by rendering the object from various distances along the z-axis, which extends outward from the image plane defined by the x- and y-axes. Each view is saved with the corresponding superellipsoid parameters used to render the object. We choose to use depth images to focus on the shape of the object, without being influenced by its appearance, such as color, texture, or lighting.

3.2 Representing Superellipsoids Into a Structural Matrix

From the rendering parameters, a structural matrix $\mathbf{M} \in \mathbb{R}^{8 \times (n\,+\,1)}$ is derived to compactly describe the object. The structural matrix consists of one structural vector m_b describing the base component, and n structural vectors m_{s_i} describing the side component i for all $i = 1, ..., n$ side components:

$$\mathbf{M} = \begin{pmatrix} m_b \ m_{s_1} \ ... \ m_{s_n} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} x_{0b} \\ y_{0b} \\ z_{0b} \\ \epsilon_{1b} \\ \epsilon_{2b} \\ a_{xb} \\ a_{yb} \\ a_{zb} \end{pmatrix} \begin{pmatrix} \mathit{\Delta} x_{s_1} \\ \mathit{\Delta} y_{s_1} \\ \mathit{\Delta} z_{s_1} \\ \epsilon_{1s_1} \\ \epsilon_{2s_1} \\ a_{xs_1} \\ a_{ys_1} \\ a_{zs_1} \end{pmatrix} \ ... \ \begin{pmatrix} \mathit{\Delta} x_{s_n} \\ \mathit{\Delta} y_{s_n} \\ \mathit{\Delta} z_{s_n} \\ \epsilon_{1s_n} \\ \epsilon_{2s_n} \\ a_{xs_n} \\ a_{ys_n} \\ a_{zs_n} \end{pmatrix} \end{pmatrix} . \qquad (2)$$

The order of side components follows a clockwise arrangement around the object's local y-axis, starting from the positive y-axis. The structural vector of the base component m_b contains its center coordinates, x_{0_b}, y_{0_b} and z_{0_b}. For each side component i, its position is represented relative to the base component as $\mathit{\Delta} x_{s_i}$, $\mathit{\Delta} y_{s_i}$ and $\mathit{\Delta} z_{s_i}$. These values indicate the distances from the center coordinates of the base component to the center coordinates of the side component along x-, y- and z-axes. We normalize all entries of the structural vectors to lie in the range $[0, 1]$, using the minimum and maximum possible values of each parameter across the dataset. For objects with less than n side components, values of the structural vectors corresponding to "non-existent" side components are set to zero. In our current setup, $n = 2$, meaning each object consists of one base component and up to two side components. The data generation process is designed to be highly flexible and can be readily extended to support more complex configurations.

3.3 Structural Encoder Model Training

Our structural encoder E_s is a convolutional neural network consisting of three convolutional layers followed by two fully connected layers. During training, we provide depth images $\mathbf{I} \in \mathbb{R}^{h \times w}$ of height h and width w. Their corresponding ground truth structural matrix $\tilde{\mathbf{M}} \in \mathbb{R}^{8 \times (n\,+\,1)}$ is constructed using the objects' rendering parameters. The model takes an image $\mathbf{I}$ as input and is trained to generate a structural matrix $\mathbf{M} = E_s(\mathbf{I})$, optimized by the mean squared error. The structural encoder is trained on 1300 objects with a single component, 5600 objects with two components and 6900 objects with three components. The numbers are chosen based on initial experiments highlighting performance discrepancies across different component counts. Each object is presented from four distances during training, enforcing a consistent structural description across views .

3.4 Description Reliability Estimation

The interpretable information encoded in $\mathbf{M}$ enables us to assess the reliability of the computed parametrization. We compare the three-dimensional superellipsoid reconstruction, represented as a point cloud $C_{\mathbf{M}}$, with the point cloud of the object depicted in the depth image $\mathbf{I}$, denoted as $C_{\mathbf{I}}$. Specifically, we compute the Chamfer distance [13] between the reconstructed point cloud $C_{\mathbf{M}}$ and the observed point cloud $C_{\mathbf{I}}$, which measures the average closest-point distance between two point sets. The resulting value $d(C_{\mathbf{M}}, C_{\mathbf{I}})$ is a reconstruction-consistency score and serves as a proxy for reliability in the parametrization. Lower values indicate closer geometric agreement and thus higher reliability of the description encoded in $\mathbf{M}$. This score is applicable regardless of whether the object is known or novel, synthetic or real.

4 Description Generation Process within a Cognitive System

This section describes how the trained structural encoder is integrated into a cognitive robotic system to generate interpretable object descriptions. We first explain the system's architecture and preprocessing steps for depth images, followed by the translation of structural matrices into human-language descriptions and schematic visualizations. Finally, we discuss how the structural matrix enables efficient transmission and low-resource operation, along with a comparison of different approaches in terms of computational and bandwidth requirements.

4.1 Model Integration into Cognitive Architecture

We integrate the trained model E_s into a cognitive robotic architecture to facilitate human-robot interaction. The architecture comprises modular components that communicate via logical message passing. Our system processes objects placed on flat surfaces such as a desk or other inspection surface. A transformation matrix, computed from the sensor's pose, aligns camera and world coordinate systems and is used to estimate the surface plane. Points are labeled as inliers if their surface normals align with the plane normal in direction and proximity. The RANSAC [18] algorithm is applied to refine the estimation further. Remaining points are clustered to segment the object point clouds. A depth image $\mathbf{I}$ is generated by projecting z-coordinates onto the x-y plane. Sparse regions are inpainted, and noise is reduced via median filtering. To improve description robustness against object placement within the depth image, we compute structural matrices for shifted variants $\mathbf{I}^{\delta}$. The matrix with the lowest Chamfer distance among all tested shifts $\delta \in \Delta$ is selected.

4.2 Human-Language Description Generation from Superellispoid Parameters

To make the output of the model easily interpretable to human users, we translate the structural matrix $\mathbf{M}$ into a set of structural predicates that capture spatial and shape-related information. During model training, $\mathbf{M}$ is computed based on base and side component assignments. For our human-language descriptions, we consider three strategies for ordering the estimated components,

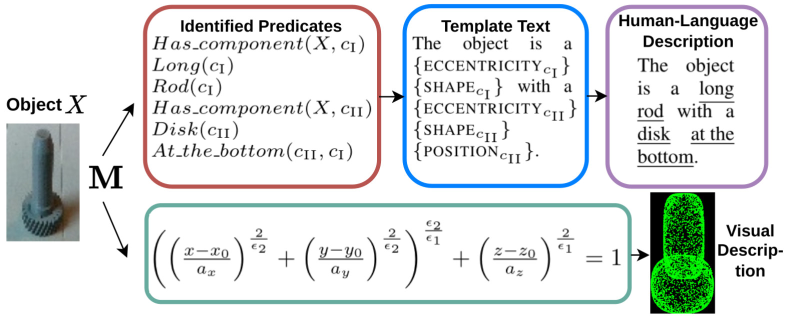

Object X
M
Identified Predicates
$Has\_component(X, c_{\mathrm{I}})$
$Long(c_{\mathrm{I}})$
$Rod(c_{\mathrm{I}})$
$Has\_component(X, c_{\mathrm{II}})$
$Disk(c_{\mathrm{II}})$
$At\_the\_bottom(c_{\mathrm{II}}, c_{\mathrm{I}})$
Template Text
The object is a {ECCENTRICITY$_{c_{\mathrm{I}}}$} {SHAPE$_{c_{\mathrm{I}}}$} with a {ECCENTRICITY$_{c_{\mathrm{II}}}$} {SHAPE$_{c_{\mathrm{II}}}$} {POSITION$_{c_{\mathrm{II}}}$}.
Human-Language Description
The object is a long rod with a disk at the bottom.
$\left(\left(\frac{x-x_0}{a_x}\right)^{\frac{2}{\epsilon_2}}+\left(\frac{y-y_0}{a_y}\right)^{\frac{2}{\epsilon_2}}\right)^{\frac{\epsilon_2}{\epsilon_1}}+\left(\frac{z-z_0}{a_z}\right)^{\frac{2}{\epsilon_1}}=1$
Visual Description

Fig. 2: An object X can be described in human language using predicates and template-based text, or visually.

since the chosen order may affect the clarity of the resulting descriptions. The three ordering strategies are:

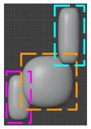
A **By component size:** The product of the main axes for component c_{I} is greater than that of component c_{II}, which in turn is greater than the product of the main axes for component c_{III},

B **By proximity to the center:** The component closest to the center of the estimated composite object is c_{I}, followed by c_{II}, and then c_{III},

C **By top-bottom-left-right arrangement:** In the x-y-plane, i.e. the depth image perspective, the component with the highest y-coordinate (topmost) and the smallest x-coordinate (leftmost) is c_{I}, followed by c_{II}, and then c_{III}.

See Tab. 1 for an example for the three options. We define the Composition predicate, $Has\_component(X, c_j)$, to indicate whether an object X includes a specific component c_j (with $j \in \mathrm{I}, \mathrm{II}, \mathrm{III}$). During model training, entries in the structural matrix $\mathbf{M}$ corresponding to absent side components are set to zero. The presence of a component c_j in an object X is inferred if the sum of its structural vector m_j exceeds an existence threshold t_{ex}. For each present component, we assign predicates describing its properties that are intuitive for human users, based on the geometric structure of the superellipsoids. For example, a *Rod* is characterized by a high edginess within one cross-sectional plane, a low edginess in the orthogonal plane and substantial elongation along its principal axis (see Technial Appenix for predicate specifications). If no specific Shape predicate such as *Disk* or *Rod* applies, we assign the generic fallback predicate *Part*. Position predicates such as On_top or To_the_left are derived from the component's spatial relations. Eccentricity predicates like *Long* are determined based on the ratio between the principal axes. Once all predicates are identified, they are used to populate gaps in predefined natural language templates, see Fig. 2. Note that our mapping strategies can be easily translated to other languages.

Table 1: Different human-language description assignments A, B and C along with the visual description for a synthetic object.

Object	A (size)	B (center proximity)	C (top-bottom-left-right)	Visual Description
	The object has a central disk with one long rod stacked on top to the right and another rod to the left at the bottom.	The object has a central disk with one rod to the left at the bottom and another long rod stacked on top to the right.	The object has a long rod with one disk to the left at the bottom and another rod further left at the bottom.	

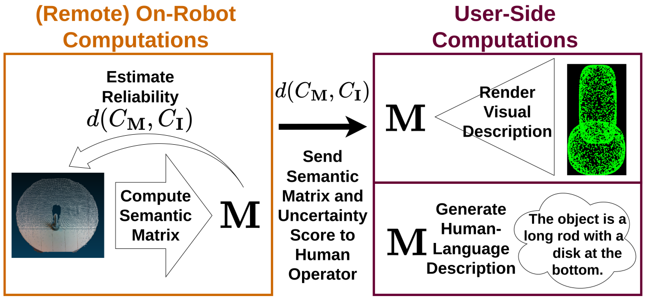

(Remote) On-Robot Computations
Estimate Reliability $d(C_{\mathbf{M}}, C_{\mathbf{I}})$
Compute Semantic Matrix
$\mathbf{M}$
$d(C_{\mathbf{M}}, C_{\mathbf{I}})$
Send Semantic Matrix and Uncertainty Score to Human Operator
User-Side Computations
$\mathbf{M}$
Render Visual Description
$\mathbf{M}$
Generate Human-Language Description
The object is a long rod with a disk at the bottom.

Fig. 3: The compact representation $\mathbf{M}$ and the reconstruction consistency score based on $d(C_{\mathbf{M}}, C_{\mathbf{I}})$ can be computed on the robot and transmitted to another device for further processing.

4.3 Superellipsoid Reconstructions as Schematic Visual Descriptions

The structural matrix $\mathbf{M}$ contains the estimated parametrizations of an object, hence we can reconstruct a 3D superellipsoid based illustration of the object. To facilitate visualization, we project the z-coordinate onto the $x-y$ plane, creating a 2D depiction of the processed object, see Fig. 2. These representations offer an alternative to lingual descriptions of object characteristics, which may be particularly beneficial for users who have difficulty comprehending human language. Conversely, for users with visual impairments, the system's human-language descriptions can be delivered via audio for users with visual impairments.

4.4 The Structural Matrix for Efficient Transmission

In low-bandwidth scenarios, the modular design of our system allows for encoding of the robot's sensor data into the compact structural matrix $\mathbf{M}$. This representation, along with the corresponding Chamfer distance $d(C_{\mathbf{M}}, C_{\mathbf{I}})$, can be transmitted effectively to a remote operator's machine (see Fig. 3). Once received, the structural data can be either translated into human-understandable descriptions or used for further analysis. The Chamfer distance provides a verifiable estimate of reliability. When the mismatch is too large, the system can abstain from producing a description, effectively signaling "I don't know".

4.5 Resource Requirements for Training, Inference and Transmission

In environments with limited bandwidth or compute resources, it becomes infeasible to access cloud-based multi-modal large language models (LLMs) or perform on-device computations. Despite being designed for edge deployment, on-device multi-modal LLMs remain relatively large. For example, Google's Gemini Nano-1 and Nano-2 have 1.8 billion and 3.25 billion parameters [21], while

DeepSeek-AI's DeepSeek-VL2-Tiny incorporates approximately 1.0 billion parameters during inference [53]. When models run directly on the robot, only the generated description needs to be transmitted.

Alternatively, compressed sensor data can be sent to a more computationally capable device for further analysis. Compression methods like DRACO [25] for point clouds or HEIF [39] for images offer high-fidelity data compression. However, the resulting data may still be too large for efficient transmission under severe bandwidth constraints. A more bandwidth-conscious solution lies in embedding-based approaches. Image encoders used in CLIP [43] or BLIP-2 [34] can encode images into compact embeddings. These representations are significantly smaller in size and can be transmitted to a more powerful device where they can be decoded into human-readable descriptions using LLMs. However, such models are computationally intensive and require extensive datasets for training. Mobile-optimized variants like the MCi0 encoder in MobileCLIP [51] reduce model size but yield opaque embeddings and depend on large datasets and strong language decoders for meaningful outputs. See Tab. 2 for a comparison of approaches.

Our structural encoder model can be efficiently trained using a single, standard graphics processing unit (GPU). Specifically, we trained the model on a single NVIDIA TITAN X (Pascal) GPU. During inference, the model can be run entirely on a central processing unit CPU. Without compression methods such as quantization or pruning, it processes a single image in 0.0398 seconds on a 12th Generation Intel Core i7-12800H. The encoded output comprises 24 values, resulting in a transmission size of 96 bytes assuming 32-bit floating-point representation (4 bytes per value).

5 Informative Value and Description Preference Assessment in a User Study

Study Design Our primary goal is to establish a robotic system that effectively communicates with a remote human user. Accordingly, we assessed the comprehensibility of our descriptions in a human user study. The study was administered online and consisted of five modules (1a, 1b, 2, 3, 4), each comprising eight single-choice questions. The remote format reflects the way users would typically interact with our system, i.e. by interpreting objects through text or images rather than through direct, physical observation. We evaluated the informative power of our human-language and visual descriptions in Modules 1a and 2, respectively, by asking participants to select the best-matching object for given descriptions. Module 1b employed GPT-generated descriptions as a strong, high-performing upper baseline against which we compared the performance of our method. Module 3 examined how well our two description variants correspond to each other, and Module 4 investigated which component order (A, B, or C) users found most useful. The first three modules (1a, 1b, 2) focused on real-world objects from two categories: (1) 3D replicas of mechanical parts used in the FetchIt! Challenge [27], and (2) everyday objects from the Yale-CMU-

Table 2: Transmission size, model complexity, training and decoding effort of different methods. Byte numbers denoted with * were measured for our user study object recordings. ✓ = interpretable, ✗ = not interpretable. N/A = not applicable. K = thousand, M = million, B = billion.

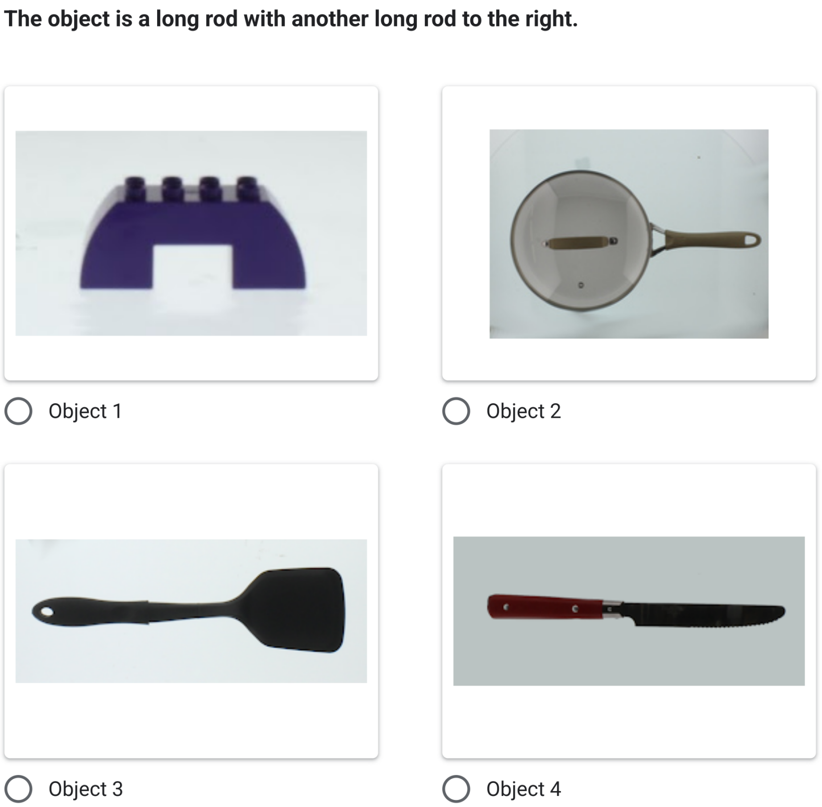
Category	**Method**	**Transmission Size (bytes, float32)**	**Model Size (# params)**	**Training Data (# samples)**	**Decoding Method**	**Interpret. Encoding**
on-device	Gemini Nano 1.0 [21]	1 per character	1.8B	undisclosed	LLM	✗
LLM	DeepSeek-VL2-Tiny [53]	(description)	3.37B / 1B		(on-robot)	
Point Cloud	Octree [15]	387 800*	N/A	N/A	N/A	✓
Compression	DRACO [25]	193 482*				
Image	WebP [24]	21 752*	N/A	N/A	N/A	✓
Compression	HEIF [39]	12 567*				
Image	ViT-B/32 [11] / CLIP [43]	2 048	87.8M	400M (real)	LLM (user-side)	✗
Encoder	ViT-L/14 [11] / BLIP-2 [34]	3 072	304M			
Only	MCi0 [51] / MobileCLIP-S0 [51]	2 048	11.4M	12M / 1B (real)		
(LLM)	MCi2 [51] / MobileCLIP-S2 [51]	2 048	35.7M			
Ours	Structural Encoder	96	6.3M	55K (synthetic)	template-based	✓

Berkeley (YCB) set [5]. Our synthetic objects were used for Modules 3 and 4. Descriptions given in Modules 1a, 2, and 3 were generated by our method, while Module 1b used descriptions generated by GPT.

Participants and Recruitment We recruited 100 participants for Modules 1a, 2, 3, and 4 and an independent sample of 50 participants for Module 1b, all via Prolific. The two samples were comparable in age, sex, and ethnicity. Mean age was 41.09±14.01 years in the 100-participant group and 42.48±14.62 years in the 50-participant group (Welch's two-sample t-test, $p = 0.58$; standardized mean difference ≈ 0.10). Sex distributions were similar (female: 56% vs 50%; Pearson's chi-square test, $p = 0.56$). Ethnicity distributions were also similar (White: 76% vs 74%; Black: 18% vs 18%; Asian: 1% vs 6%; Mixed: 3% vs 2%; Other: 1% vs 0%; Pearson's chi-square test, $p = 0.44$). Given that the task involved reading text descriptions in English, the recruitment criteria included being a native English speaker, having no history of dyslexia or literacy-related impairments, not being diagnosed with ADHD, and having normal vision or vision corrected to normal levels (e.g., with glasses, contact lenses, ...). Participants completed the study remotely on their own tablets or desktops. The study was reviewed and approved as Exempt under Exempt Category 3(i)(A) by our Institutional Review Board. Each participant provided informed consent and was compensated at a rate of $11.22 per hour. Based on their responses, we identified key insights aligned with our research questions (RQs).

Metrics We quantify the results of RQ 1- 3 using two metrics. *Per-question description validity* measures how effectively each description guided participants to the corresonding object. It is calculated as the proportion of participants who selected the target object (i.e., the one from which the description was generated) for a given question. *Per-participant description validity* captures how well individual participants were able to identify the target objects based on the provided description. It is calculated as the number of times a participant selected the target objects divided by the total number of questions they answered. Both metrics are expressed as percentages, with higher values indicating that more participants identified the intended answer. Participant preference for each component order (RQ4) is quantified using the proportion of participants who selected a given description as most helpful for a question.

Module 1a - RQ1a: Do our human-language descriptions support object identification?
We hypothesize that our method generates linguistic descriptions that effectively communicate object characteristics for human understanding, such that participants are able to correctly identify the described objects.
Methods: Participants were given a description of an object along four images of objects, only one of which was used to generate the description. They were asked to select the image they believed best matched the description (see Fig. 4 for an example).
Results: *The strong agreement between participant responses and our intended description-object pairs shows that our method successfully supports object identification.* Across all eight single-choice questions, 88% of participants selected



The object is a long rod with another long rod to the right.

○ Object 1

○ Object 2

○ Object 3

○ Object 4

Fig. 4: Example question of Module 1a, where participants are shown a text description generated by our method and four object images, only one of which was used to generate the description, and asked to select the best match.

the object from which the description was generated (see per-question description validity for RQ1a in Fig. 5). The median (P_{50}) per-participant description validity was 87.5%, implying that at least half of the participants chose the object corresponding to the given text description for 7 out of 8 questions. *Some object descriptions led to more varied interpretations.* The spread of per-participant description validity values (5th percentile = 62.5%, 20th percentile = 75%, see Fig. 5), suggests that while many participants aligned with the assumed best match, for others some of our descriptions did not resonate as strongly. We suspect that disagreements arise from differences in how participants mentally categorize object parts. For example, in Question 7_{RQ1}, a toy drill's handle was described as a "rod", which may have conflicted with participants' expectations, while the same term was widely considered appropriate for a large gear in Question 1_{RQ1}.

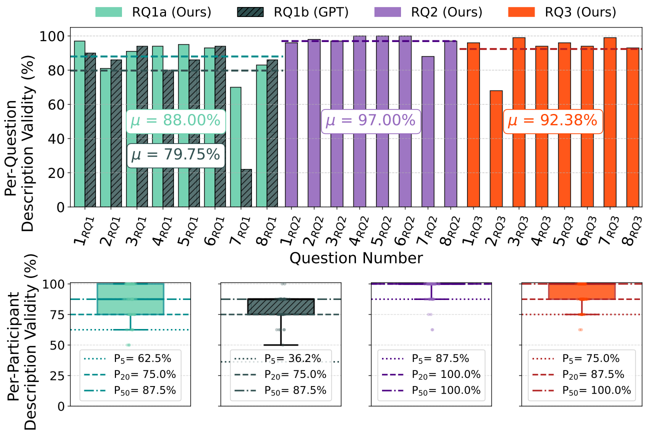


Fig. 5: **Top:** Per-question validity, shown as bars with mean (μ) values for each question. **Bottom:** Per-participant validity, shown as boxplots with 5th, 20th, and 50th percentiles (P_5, P_{20}, P_{50}). Boxes indicate the interquartile range (25th–75th percentiles), whiskers extend to the most extreme observed values within the 5th–95th percentile range, and individual participant scores are overlaid as jittered points. Overall, the provided descriptions strongly guided participants to the intended object-description pairs.

Module 1b - RQ1b: How do our descriptions compare to GPT-generated descriptions in terms of object identification effectiveness? We hypothesize that GPT-generated descriptions, given its resource-intensive infrastructure, will provide an approximate upper reference point for object identification performance. By comparing our method to GPT, we aim to contextualize the effectiveness of our compact, fixed-vocabulary approach.
Methods: We used the same objects and experimental structure as in Module 1a, but replaced our method's descriptions with those generated by the OpenAI GPT-4o model [41]. The model is accessed through the OpenAI API, the temperature is set to 0.2 and all other hyperparameters to their default values. We prompted GPT to generate simple, human-readable descriptions, specifying parts and their spatial relationships while restricting labels or object-specific terms. Without these constraints, GPT defaulted to familiar object labels (e.g., ”power drill”) and related semantic terms (e.g., ”drill bit,” ”chuck”), even though such associations are not applicable for novel objects.

Results: *Our method performs comparably to GPT in enabling accurate object identification.* Both description methods resulted in strong alignment between participant selections and intended matches. Across all eight questions, average description validity remained high, with at least half of participants correctly identifying the object in 7 out of 8 cases for both methods (P_{50} = 87.5%, see

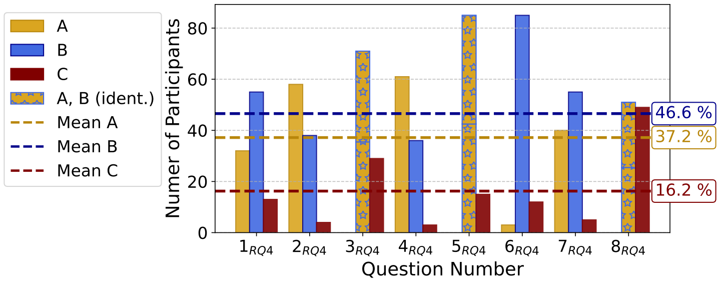


Fig. 6: Participants favored component order assignment B, followed by A, with C receiving the least preference. The patterned bars represent instances where order assignments A and B resulted in identical descriptions.

Fig. 5). A Mann–Whitney U test indicated that participants were slightly more likely to select the intended object when guided by our descriptions ($U = 3480$, $p < 0.001$, $r = 0.39$), consistent with Welch's t-test ($t = 3.56$, $p = 0.0006$, Cohen's $d = 0.64$). *Each method appeared more effective for certain objects.* Three objects were more frequently identified when described by GPT, four were more frequently identified when described by our method. *One object was not identifiable for many participants (73 %).* The small gear (Question 7_{RQ1}), was described by GPT using the phrase "cone", which may have been unintuitive for participants. Our method described the same object as a composite of "disks", which may have provided clearer cues, though this object yielded the lowest description validity across both methods.

Module 2 - RQ2: Do our visual descriptions allow object identification?

We hypothesize that our method generates visual descriptions that effectively communicate object characteristics for human understanding, such that participants are able to correctly identify the corresponding objects.

Methods: In this module, participants completed two types of matching tasks. For the first type, they were shown an image of an object along with four visual descriptions, only one of which was generated from the presented object. For questions of the second type, they were shown a visual description alongside four images of objects. Only one of the depicted objects was used to generate the given visual description. Participants were asked to identify the best matching description-object pairs.

Results: *The strong alignment between participant selections and our intended object-description pairs indicates that our visual descriptions effectively convey object characteristics*, with an average of 97% participants selecting the intented answeres across all questions (see per-question description validity for RQ2 in Fig. 5). *Participants seemed to find the visual descriptions at least as informative as the human language descriptions,* indicated by higher overall per-question description validity and higher per-participant description validity percentiles for

Module 2 compared to Module 1a. We suspect that our visual descriptions not only clearly represent objects but also reduce cognitive load. Unlike text, they require less mental effort to interpret and are not dependent on specific linguistic terms, which may vary in how intuitively they are understood.

Module 3 - RQ3: Do our text descriptions align with the underlying superellipsoid-based representation (visual description) of the object?
We hypothesize that our visual descriptions will convey object meaning consistently with the corresponding text descriptions, such that they can effectively replace or complement language in promoting more inclusive interaction.
Methods: Participants were presented with a text description and four visual descriptions, only one of which corresponds to the same object as the text description. They were asked to select the visual description that best fits the given text description.
Results: *The visual descriptions align generally well with the human language descriptions*, as indicated by 92.38% of participants selecting the visual description belonging to the same object as the given text description, see per-question description validity for RQ3 in Fig. 5. The median per-participant description validity shows that at least half of the participants found the corresponding description pairs to be the most suitable (P_{50} for RQ3 in Fig. 5). For Question 2_{RQ3}, only 68% of participants selected the intended pair, as some chose visual descriptions that matched the objects' shapes but had mismatched spatial arrangements.

Module 4 - RQ4: Which component order (A, B, C) do users find most useful?
We aim to investigate participants' preferences for the order of components used in generating our human-language descriptions.
Methods: The component orders (see Sec. 4.2) are based on size (A), proximity to the center (B), or a top-bottom-left-right arrangement (C). Participants were shown an image of an object alongside text descriptions generated using A, B and C, presented in a random sequence. They were asked to choose the description they considered most helpful for identifying the object. For three questions, A and B resulted in identical descriptions. In such cases, participants had only two descriptions to choose from: the A, B description and C.
Results: *Component order* B *(proximity to center), is the overall preferred option.* It was selected in 46.6% of responses and chosen most often in three of eight questions (Fig. 6). Size-based order A, is the second preferred option and ranked highest in two questions, while top-bottom-left-right component order C is selected the least often. Participant preferences showed varying levels of consensus across questions. Certain questions, such as Question 5_{RQ4} and 6_{RQ4}, show a high level of agreement among participants, whereas others, like 1_{RQ4} and 8_{RQ4}, reveal a more evenly distributed preference, suggesting less consensus. We assume that participants generally prefer descriptions that reference centrally located components, as these are most likely more intuitive.

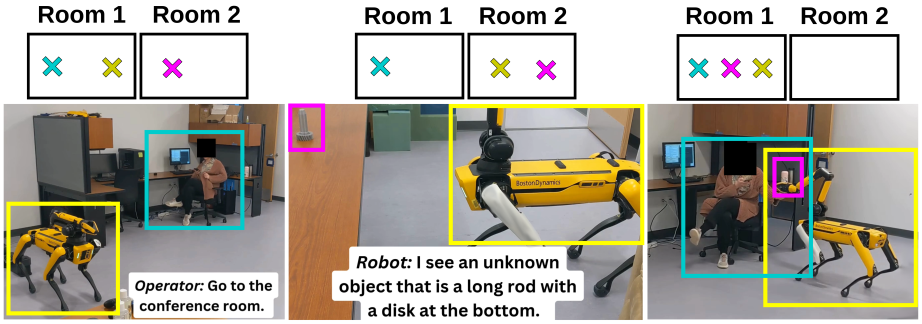

Room 1
Room 2
Room 1
Room 2
Room 1
Room 2
Operator: Go to the conference room.
Robot: I see an unknown object that is a long rod with a disk at the bottom.
Boston Dynamics

Fig. 7: The operator and the robot (blue and yellow marker) start in Room 1. The operator instructs the robot to move to Room 2 (conference room) to describe an unknown object (pink marker). The robot then retrieves the object.

6 On-Robot Task Demonstration

We showcase our method in a real-world scenario involving a human operator, a robot and an unknown object. The object is a 3D replica from the FetchIt! Challenge [27], placed on a table. We use the Boston Dynamics Spot robot, equipped with a gripping arm and a depth sensor. A smartphone with an Automatic Speech Recognition (ASR) provides natural language input, while the cognitive architecture runs on a Linux device, facilitating communication between devices. The robot is initialized with various locations (e.g. “conference room”) but has no knowledge of the objects it will be asked to retrieve. Initially, both the human operator and the robot are located in the same room (Room 1 in Fig. 7). The operator instructs the robot to navigate to the conference room (Room 2 in Fig. 7). The robot proceeds to the designated room via its navigation graph, while the operator remains in the initial room. Upon arrival, the robot notifies the operator. The operator asks "*What do you see?*" and the robot enters a predetermined observation pose.

The robot groups point cloud data into object clusters. If any of the object clusters is not recognized under pre-existing categories of YOLOv5 [50] loaded in the cognitive architecture, the object is assigned to be “unknown” and described by our system. The structural encoder processes the depth image derived from the unknown object’s point cloud to compute the structural matrix. These values are translated into our defined structural predicates and passed into the cognitive architecture’s natural language generation pipeline. A complete predicate representation of the observation is provided to an LLM [26], which generates the full response: : “*I see an unknown object that is a long rod with a disk at the bottom.*” The operator instructs the robot to deliver the unknown object to a pre-defined drop-off location. This goal is submitted to the cognitive architecture, which utilizes its planner [29] to generate a sequence of actions to achieve the task. The robot grabs the object, goes to the drop off location in Room 1, and hands the object to the operator.

7 Conclusion

We present a method for describing novel objects using geometric mereological features that balances structural interpretability, computational efficiency, and communicative usefulness within a constrained representational scope. By relying exclusively on synthetically generated data, the approach allows precise control over object structure and eliminates the risk of spurious samples or the need for manual annotation. An efficient structural encoder extracts a compact and interpretable set of geometric parameters describing component shape, eccentricity, and spatial relations. By design, the encoder operates with modest computational resources, trading unrestricted expressiveness for reliability and deployability. The resulting object representation is highly compressed, enabling low-bandwidth transmission. Reconstruction-based comparison between perceived objects and their reconstructed versions further provides an internal reliability estimate. Within a cognitive architecture, this representation supports both linguistic and visual description generation, enabling consistent object communication across modalities.

Rather than conducting a large-scale evaluation on novel objects, which is beyond the scope of this paper, we highlight the system's potential through a user-centered evaluation, which we consider essential for assessing the practical value of our method. Participants reliably matched generated descriptions to their corresponding objects, achieving object identification performance comparable to GPT-based descriptions despite the fixed vocabulary and substantially lower computational requirements of our approach. GPT was used as an exploratory upper bound rather than a competitive baseline, to contextualize the expressive capacity of a lightweight description system.

Comparison of different description generation strategies showed that ordering components by proximity to the object center was consistently preferred by users. The real-world feasibility of our method is demonstrated through a robotic scenario, in which an unknown object was successfully described to a human operator.

While these results support the core assumptions of the framework, they represent an initial step toward broader generalization.

8 Limitations and Future Work

While our method demonstrates promising results for generating interpretable descriptions of novel objects, several limitations point to directions for future improvement.

Model performance was found to vary with object position in the depth image. While positional shifts helped stabilize results, the underlying issue stems from the assumption of a centrally positioned base component. Future work will explore different component assignment strategies to enable more robust handling of objects.

While our study shows that visual descriptions are generally more effective than text (Module 1a vs. 2), the trade-offs between these modalities, such as cognitive load or accessibility, have not yet been fully explored.

While users consistently preferred ordering components by proximity to the object center, the underlying reasons for this preference remain unclear. Factors such as reduced cognitive load, alignment with spatial intuitions, or more efficient visual search may contribute to this effect. A targeted user study could explicitly investigate these factors, for example by measuring task completion time, error rates, and subjective workload.

The current system supports objects with up to three components. Consequently, objects with more complex structures are represented as simplified approximations. Although participants could still identify simplified objects accurately, extending the geometric representation to handle additional components and hierarchical decompositions will broaden applicability.

Real-world inputs are often subject to sensor noise, occlusions, and imperfect depth measurements. Approaches such as data augmentation with synthetic noise and noise-robust fitting techniques could be considered to increase resilience.

While the system computes a reliability score for object descriptions, this information is not yet conveyed to users. Including confidence indicators or expressions of uncertainty (e.g., "This object is likely a rod with a disk") could enhance transparency and user trust.

Furthermore, ategorizing a continuous world into discrete labels reflects the inherent vagueness and context dependence of human language [35]. While terms like "cylinder" offer general shape categories, they may not be intuitive for all users. To improve clarity, we discretized superellipsoids into more familiar terms. However, labels like "rod" may fit some contexts (e.g., screws) but not others, and some shapes lie between categories. In future work, we aim to assess the degree to which a predicate applies. Additionally, user studies on label interpretability and context suitability could help to refine our predicate assignments.

References

1. Alaniz, S., Mancini, M., Akata, Z.: Iterative superquadric recomposition of 3d objects from multiple views. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18013–18023 (October 2023). https://doi.org/10.48550/arXiv.2309.02102
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). p. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2019). https://doi.org/10.65109/KCZB5817
3. Barr, A.H.: Superquadrics and angle-preserving transformations. IEEE Computer Graphics and Applications **1**(1), 11–23 (1981). https://doi.org/10.1109/MCG.1981.1673799

4. Blender Online Community: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org
5. Calli, B., Singh, A., Bruce, J., Walsman, A., Konolige, K., Srinivasa, S., Abbeel, P., Dollar, A.M.: Yale-cmu-berkeley dataset for robotic manipulation research. The International Journal of Robotics Research **36**(3), 261–268 (2017). https://doi.org/10.1177/0278364917700714
6. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: SpatialRGPT: Grounded spatial reasoning in vision-language models (2024), https://openreview.net/forum?id=JKEIYQUSUc
7. Chiou, M., Epsimos, G.T., Nikolaou, G., Pappas, P., Petousakis, G., Mühl, S., Stolkin, R.: Robot-assisted nuclear disaster response: Report and insights from a field exercise. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4545–4552 (2022). https://doi.org/10.1109/IROS47612.2022.9981881
8. Chitikena, H., Sanfilippo, F., Ma, S.: Robotics in search and rescue (sar) operations: An ethical and design perspective framework for response phase. Applied Sciences **13**(3) (2023). https://doi.org/10.3390/app13031800, https://www.mdpi.com/2076-3417/13/3/1800
9. Cruz, S., Doctor, K., Funk, C., Scheirer, W.: Open issues in open world learning. AAAI AI Magazine **46**(2) (2025). https://doi.org/10.1002/aaai.70001
10. Damaševičius, R., Bacanin, N., Misra, S.: From sensors to safety: Internet of emergency services (ioes) for emergency response and disaster management. Journal of Sensor and Actuator Networks **12**(3) (2023). https://doi.org/10.3390/jsan12030041, https://www.mdpi.com/2224-2708/12/3/41
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=YicbFdNTTy
12. Drew, D.: Multi-agent systems for search and rescue applications. Current Robotics Reports **2** (03 2021). https://doi.org/10.1007/s43154-021-00048-3
13. Dubuisson, M.P., Jain, A.: A modified hausdorff distance for object matching. In: Proceedings of 12th International Conference on Pattern Recognition. vol. 1, pp. 566–568 vol.1 (1994). https://doi.org/10.1109/ICPR.1994.576361
14. Elmoznino, E., Jiralerspong, T., Bengio, Y., Lajoie, G.: Towards a formal theory of representational compositionality. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=fXCfT7ErvL
15. Elseberg, J., Borrmann, D., Nüchter, A.: A comparison of nearest-neighbor-search algorithms and implementations for efficient shape registration. In: Journal of Photogrammetry and Remote Sensing (ISPRS). vol. 66, pp. 208–216. Elsevier (2011)
16. Fedele, E., Sun, B., Guibas, L., Pollefeys, M., Engelmann, F.: SuperDec: 3D Scene Decomposition with Superquadric Primitives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025). https://doi.org/10.48550/arXiv.2504.00992
17. Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K., et al.: Foundation models in robotics: Applications, challenges, and the future. International Journal of Robotics Research (2025). https://doi.org/10.1177/02783649241281508

18. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography pp. 726–740 (1987). https://doi.org/10.1016/B978-0-08-051581-6.50070-2, https://www.sciencedirect.com/science/article/pii/B9780080515816500702
19. Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., Sadigh, D.: Physically grounded vision-language models for robotic manipulation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 12462–12469 (2024). https://doi.org/10.1109/ICRA57147.2024.10610090
20. Gao, Z., Yi, R., Huang, Y., Chen, W., Zhu, C., Xu, K.: Learning part-aware 3d representations by fusing 2d gaussians and superquadrics. CoRR **abs/2408.10789** (2024), https://doi.org/10.48550/arXiv.2408.10789
21. Gemini Team Google: Gemini: A family of highly capable multimodal models (2024). https://doi.org/10.48550/arXiv.2312.11805
22. Goel, S., Lymperopoulos, P., Thielstrom, R., Krause, E., Feeney, P., Lorang, P., Schneider, S., Wei, Y., Kildebeck, E., Goss, S., Hughes, M.C., Liu, L., Sinapov, J., Scheutz, M.: A neurosymbolic cognitive architecture framework for handling novelties in open worlds. Artificial Intelligence **331**, 104111 (2024). https://doi.org/10.1016/j.artint.2024.104111, https://www.sciencedirect.com/science/article/pii/S000437022400047X
23. Goodwill, J., Wilson, C., MacKinnon, J.: Current ai technology in space. In: Krittanawong, C. (ed.) Precision Medicine for Long and Safe Permanence of Humans in Space, pp. 239–250. Academic Press (2025). https://doi.org/10.1016/B978-0-443-22259-7.00006-0, https://www.sciencedirect.com/science/article/pii/B9780443222597000060
24. Google: Webp compression. https://developers.google.com/speed/webp (2010), accessed: 2025-04-11
25. Google: Draco: 3d data compression. https://github.com/google/draco (2017), accessed: 2025-04-11
26. Grattafiori, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
27. Han, Z., Allspaw, J., LeMasurier, G., Parrillo, J., Giger, D., Ahmadzadeh, S.R., Yanco, H.A.: Towards mobile multi-task manipulation in a confined and integrated environment with irregular objects. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). vol. 3, p. 11025–11031. IEEE (May 2020). https://doi.org/10.1109/icra40945.2020.9197395, http://dx.doi.org/10.1109/ICRA40945.2020.9197395
28. Hancock, P., Billings, D., Schaefer, K., Chen, J., de Visser, E., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Human factors **53**, 517–27 (10 2011). https://doi.org/10.1177/0018720811417254
29. Hoffmann, J., Nebel, B.: The FF planning system: Fast plan generation through heuristic search. Journal of Artificial Intelligence Research **14**, 253–302 (2001)
30. Holder, L., Langley, P., Loyall, B., Senator, T.: Introduction to open-world ai. Artificial Intelligence p. 104393 (2025). https://doi.org/10.1016/j.artint.2025.104393, https://www.sciencedirect.com/science/article/pii/S0004370225001122
31. Hu, Y., Xie, Q., Jain, V., Francis, J., Patrikar, J., Keetha, N., Kim, S., Xie, Y., Zhang, T., Zhao, S., Chong, Y.Q., Wang, C., Sycara, K., Johnson-Roberson, M., Batra, D., Wang, X., Scherer, S., Kira, Z., Xia, F., Bisk, Y.: Toward general-purpose robots via foundation models: A survey and meta-analysis (2023). https://doi.org/10.48550/arXiv.2312.08782

32. Jiang, S., Zhao, Q., Rahmani, H., Soh, D.W., Liu, J., Zhao, N.: Gaussianblock: Building part-aware compositional and editable 3d scene by primitives and gaussians. CoRR **abs/2410.01535** (2024), https://doi.org/10.48550/arXiv.2410.01535
33. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 119, pp. 5338–5348. PMLR (13–18 Jul 2020). https://doi.org/10.48550/arXiv.2007.04612, https://proceedings.mlr.press/v119/koh20a.html
34. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023). https://doi.org/10.48550/arXiv.2301.12597
35. Lim, W., Wu, Q.: Vague language and context dependence. Frontiers in Behavioral Economics (2023). https://doi.org/10.3389/frbhe.2023.1014233, https://www.frontiersin.org/articles/10.3389/frbhe.2023.1014233/full
36. Lin, C., Fan, T., Wang, W., Nießner, M.: Modeling 3d shapes by reinforcement learning. In: European Conference on Computer Vision (ECCV). pp. 545–561. Springer (2020). https://doi.org/10.1007/978-3-030-58607-2_32
37. Ma, T., Wang, Z., Zhou, J., Wang, M., Liang, J.: Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping (2024), https://arxiv.org/abs/2411.12286
38. Marr, D.: Vision: a computational investigation into the human representation and processing of visual information. Vision: a computational investigation into the human representation and processing of visual information. (1982). https://doi.org/10.1016/0022-2496(83)90030-5
39. MPEG: High efficiency image file format (heif). https://mpeg.chiariglione.org/standards/mpeg-h/image-file-format (2015), accessed: 2025-04-11
40. Ommer, B., Buhmann, J.M.: Learning the compositional nature of visual object categories for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010). https://doi.org/10.1109/tpami.2009.22
41. OpenAI: Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/ (2024), accessed: 2025-04-14
42. Qiu, Z., Zhao, H., Wang, S.: Applications and challenges of artificial intelligence in aerospace engineering. In: 2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD). pp. 970–974 (2023). https://doi.org/10.1109/ICAIBD57115.2023.10206205
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021). https://doi.org/10.48550/arXiv.2103.00020, https://api.semanticscholar.org/CorpusID:231591445
44. Reddy, P., Guerrero, P., Mitra, N.J.: Search for concepts: Discovering visual concepts using direct optimization (2022). https://doi.org/10.48550/ARXIV.2210.14808, https://arxiv.org/abs/2210.14808
45. Shen, W., Wei, Z., Huang, S., Zhang, B., Fan, J., Zhao, P., Zhang, Q.: Interpretable compositional convolutional neural networks. In: International Joint Conference on Artificial Intelligence (IJCAI) (2021). https://doi.org/10.24963/ijcai.2021/409

46. Si, Z., Zhu, S.C.: Learning and-or templates for object recognition and detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013). https://doi.org/10.1109/TPAMI.2013.35
47. Sinha, S., Premsri, T., Kordjamshidi, P.: A survey on compositional learning of AI models: Theoretical and experimental practices (2024), https://openreview.net/forum?id=BXDxwItNqQ, survey Certification
48. Stone, A., Wang, H., Stark, M., Liu, Y., Phoenix, D.S., George, D.: Teaching compositionality to cnns. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2017-January (2017). https://doi.org/10.1109/CVPR.2017.85
49. Surmann, H., Daun, K., Schnaubelt, M., Von Stryk, O., Patchou, M., Böcker, S., Wietfeld, C., Quenzel, J., Schleich, D., Behnke, S., Grafe, R., Heidemann, N., Slomma, D., Kruijff-Korbayova, I.: Lessons from robot-assisted disaster response deployments by the german rescue robotics center task force. Journal of Field Robotics **41** (12 2023). https://doi.org/10.1002/rob.22275
50. Ultralytics: YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com (2021)
51. Vasu, P., Pouransari, H., Faghri, F., Vemulapalli, R., Tuzel, O.: Mobileclip: Fast image-text models through multi-modal reinforced training. pp. 15963–15974 (06 2024). https://doi.org/10.1109/CVPR52733.2024.01511
52. Whorf, B.L.: Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. MIT Press (1956)
53. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., Ruan, C.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding (2024), https://arxiv.org/abs/2412.10302
54. Yeh, C.K., Kim, B., Arik, S., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 20554–20565. Curran Associates, Inc. (2020). https://doi.org/10.48550/arXiv.1910.07969
55. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction in robotics. In: 8th Annual Conference on Robot Learning (CoRL) (2024), https://openreview.net/forum?id=GVX6jpZOhU
56. Zeng, F., Gan, W., Wang, Y., Liu, N., Yu, P.S.: Large language models for robotics: A survey (2023). https://doi.org/10.48550/arXiv.2311.07226
57. Zhu, L., Chen, Y., Torralba, A., Freeman, W., Yuille, A.: Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2010). https://doi.org/10.1109/CVPR.2010.5539865