

Towards Social Foundation Models: A Framework and Synthetic Dataset for Grounding Visual Perspective Taking in Robots

Joel Currie ^{*†}
joel.currie@iit.it

Enrico Piacenti ^{*}
enrico.piacenti@iit.it

Gioele Migno ^{*}
gioele.migno@iit.it

Mohammad Gharb ^{*}
mohammad.gharb@iit.it

Davide De Tommaso ^{*}
davide.detommaso@iit.it

Agnieszka Wykowska ^{*}
agnieszka.wykowska@iit.it

Abstract

The next frontier in robotics is the creation of truly collaborative agents that share our physical space. Achieving this requires robots to develop foundational socio-cognitive abilities. One of them is the ability to establish shared spatial representations between interacting agents. While modern Vision-Language Models possess powerful semantic capabilities, their understanding is not grounded in metric space, preventing them from mastering core skills such as Visual Perspective Taking, the ability to understand the world from another agent’s viewpoint. We argue this is not an architectural limitation but a data problem, leading us to propose *Social Foundation Models*: a new class of models designed to master a curriculum of foundational socio-cognitive primitives. This paper presents a foundational blueprint for this vision. We reformulate Visual Perspective Taking as a 6-DOF pose regression task and introduce SynthVPT, a large-scale, synthetic dataset of procedurally generated RGB images with precise ground-truth annotations. We then present a conceptual framework to exploit the rich prior knowledge of a pre-trained Vision-Language Model, fine-tuning it on our data to create a general-purpose spatial reasoner. This methodology provides a tangible and scalable pathway for teaching embodied agents the foundational socio-cognitive skills needed for genuine human-robot collaboration, moving beyond simple instruction-following towards a truly shared reality.

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**; **Computer vision**; **Spatial and physical reasoning**; • **Computer systems organization** → **Robotics**; • **Human-centered computing** → **Human-robot interaction**.

Keywords

Social Foundation Models, Visual Perspective Taking, Visual Language Models, Visual Grounding, Spatial Reasoning, Synthetic Data, Embodied-AI, Human-Robot Interaction

1 Introduction

Foundation models, large-scale neural networks trained on vast and diverse datasets, have recently emerged as a transformative paradigm in Artificial Intelligence (AI) [7]. This paradigm is accelerating advancements in robotics, where foundation models have demonstrated remarkable success in domains such as manipulation,

navigation, and planning [19, 35, 50, 58]. However, the success of these robotics-focused models stems from their training on extensive datasets of embodied, task-specific data (e.g., robot trajectories, grasp poses) [8]. In contrast, the broader class of general-purpose Vision-Language Models (VLMs), trained on web-scale data, excel at semantic understanding but inherently lack the precise, metric data needed for grounded spatial reasoning. Concurrently, a growing body of work shows that the socio-cognitive capacities of these powerful semantic models are brittle and far from robust [9, 12, 29, 51, 54]. We argue that these two limitations are deeply connected: the deficit in task-specific, geometric data is a critical barrier preventing these models from developing the nuanced collaborative reasoning necessary for successful Human-Robot Interaction (HRI) [17].

We therefore propose *Social Foundation Models*: a new class of foundation models fine-tuned on data that enables embodied agents like robots to master a curriculum of foundational socio-cognitive primitives. The primary goal is to build complex social reasoning from these simple, reusable building blocks (e.g., action prediction, intention recognition). The cornerstone of this predictive ability is mastering a foundational cognitive skill: Visual Perspective Taking (VPT) [14, 15, 17, 18]. This capability is the essential first step towards shared understanding, allowing a robot to move beyond an egocentric view and interpret human instructions and intentions within a shared context. It is the critical upstream skill that in humans enables crucial downstream social tasks, including joint action [21], social navigation [37] and mental/affective/goal state inference [6, 22, 44].

Consider a collaborative manufacturing setting: a robot and a human are looking at a complex component from opposite sides. The human asks for “the bolt on the left,” but from the robot’s perspective, that same bolt is on the right. The robot must mentally adopt the human’s viewpoint to identify the correct part. This need extends to social navigation, where an autonomous vehicle must infer a pedestrian’s visual field to predict their intent. In both scenarios, the robot must reason about spatial relationships from distinct viewpoints (its own and its collaborator’s) by mapping between diverging frames of reference.

Existing VPT solutions in robotics often rely on explicit geometric modelling as a prerequisite for object identification and pose estimation [20, 34, 43] or rule-based pipelines [18, 56], which, while effective in constrained environments, lack the flexibility and generalisability for real-world HRI. While modern VLMs offer this flexibility, their impressive performance in general scene understanding [40] does not extend to precise spatial reasoning. A growing body of work shows they struggle to infer exact object poses and relative orientations [23, 24, 53]. This deficit is not seen as an architectural

^{*} Social Cognition in Human-Robot Interaction Unit, Italian Institute of Technology, Genoa, Italy

[†] University of Aberdeen, Aberdeen, United Kingdom

limitation, but rather a consequence of their training on web-scale data that lacks explicit, grounded geometric information [10, 17, 41].

Simulated environments offer a promising solution to this data gap, as they allow for the trivial generation of scalable datasets with perfect, noise-free ground truth for structured spatial relationships [52]. Recent work has validated this approach, showing that fine-tuning VLMs with synthetic spatial data can improve VPT and spatial reasoning more broadly [48]. However, the authors’ choice of using the question-answer (QA) format limited the model to simple allocentric left-right judgements, a level of abstraction unsuitable for many embodied HRI tasks that require precise metric understanding. Notably, the study confirmed that noiseless synthetic annotations were more effective than pseudo-annotating real images and that the learned skills could generalise to real-world images, demonstrating both the potential of the synthetic data approach and the need for a more geometrically grounded methodology.

To address the key limitation preventing VLMs from performing VPT, we propose a fine-tuning framework built upon a novel, large-scale synthetic dataset. Current models, despite their powerful visual representations, lack the grounded data needed for metric spatial reasoning. Our proposed approach aims to remedy this by using a procedurally generated dataset of geometrically precise renders to teach a pre-trained VLM how to perform 6-DOF pose estimation of a target agent’s viewpoint. By training the model on this targeted data, we aim to transform it from a passive scene describer into an active spatial reasoner. This work seeks to demonstrate that the rich visual features inherent in pre-trained VLMs can be effectively channelled to solve complex, embodied AI tasks, and in doing so, establish a scalable blueprint for teaching grounded socio-cognitive skills.

Our contributions are as follows:

- A conceptual framework for developing *Social Foundation Models* grounded in metric spatial reasoning through training on large-scale synthetic data, with the core socio-cognitive skill of VPT serving as a key case study.
- An open-source proof-of-concept synthetic dataset and generation pipeline built in NVIDIA Omniverse, providing RGB images paired with annotated ground-truth 4×4 transformation matrices for supervised learning.

This work supports the future development of spatially aware, embodied robots capable of interpreting perspective-dependent instructions, reasoning about what their human partners can and cannot see, and ultimately engaging in safer, more fluid collaboration. By providing a conceptual methodology for teaching a foundational socio-cognitive skill like VPT, we lay the necessary groundwork for the future development of the *Social Foundation Models* we propose. This moves artificial agents beyond simple task execution and towards a genuine shared understanding within human-centric environments.

2 Method

To advance toward *Social Foundation Models*, we propose a conceptual framework that enables VLMs to acquire embodied metric spatial reasoning, a prerequisite for our chosen socio-cognitive task of VPT. This framework is instantiated through our primary

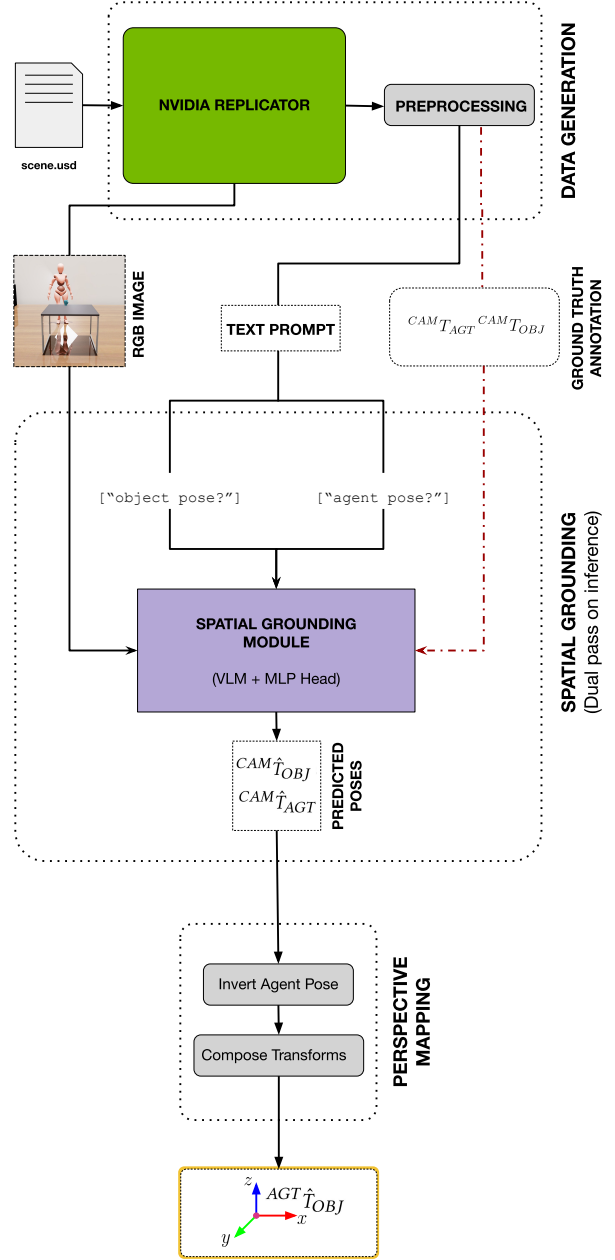


Figure 1: An overview of our proposed framework’s architecture. The diagram illustrates two distinct processes: the offline training setup $-\cdot-$, and the online inference pipeline $-$. The offline stage uses NVIDIA Replicator and a ‘Preprocessing’ step to generate the ground-truth dataset required to train the ‘Spatial Grounding Module’. During inference, this trained module takes a sample image and in two separate queries predicts the camera-relative poses of the object (${}^{CAM}\hat{T}_{OBJ}$) and agent (${}^{CAM}\hat{T}_{AGT}$). These predictions are then deterministically transformed by the ‘Perspective Mapping’ stage to compute the final 6-DOF agent-relative pose, ${}^{AGT}\hat{T}_{OBJ}$.

contribution: a large-scale synthetic dataset that provides geometrically precise, fully annotated 6-DOF spatial data. Building upon this dataset, we operationalise the framework (see Figure 1) through a three-stage pipeline:

- i. **Dataset Generation:** We first describe the completed synthetic data generation pipeline, implemented in NVIDIA Omniverse and Replicator, which produces large volumes of high-fidelity images with exact ground-truth pose annotations. This dataset provides the embodied spatial structure required for metric reasoning tasks, with the HRI context specific to VPT.
- ii. **Spatial Grounding Module (SGM):** The SGM forms the core component of our proposed framework. A pre-trained VLM is to be extended with a lightweight regression head and fine-tuned on the dataset to predict two key quantities from a single RGB image: the object pose (${}^{CAM}\hat{T}_{OBJ}$) and the agent pose (${}^{CAM}\hat{T}_{AGT}$), both expressed in the camera reference frame. Through this process, the model learns to ground its visual representations in physically consistent spatial coordinates.
- iii. **Perspective Mapping:** The final stage deterministically combines the two predicted poses to infer the object’s pose from the agent’s viewpoint (${}^{AGT}\hat{T}_{OBJ}$). This step completes the perspective-taking transformation, translating the model’s camera-centric understanding into an agent-centric one.

This three-stage framework offers a generalisable method for endowing VLMs with embodied spatial reasoning capabilities through its first two stages: *Dataset Generation* and *Spatial Grounding*. These components provide a flexible foundation, applicable for training models to perform a wide range of embodied reasoning tasks. The third stage, *Perspective Mapping*, is specifically tailored to VPT, as it transforms spatial representations into an agent-centric reference frame. By explicitly grounding visual representations in metric space, the framework establishes a foundation upon which future *Social Foundation Models* can build richer, more adaptive forms of human–robot understanding.

2.1 Dataset

2.1.1 Scene Design

Our dataset is built within the NVIDIA Omniverse ecosystem [2], with its seed being a single, high-fidelity base scene designed to isolate the core socio-cognitive skill of VPT in a controlled, simplified task. This scene was constructed in NVIDIA Isaac Sim and depicts a minimalist indoor environment containing a table, a target object, and a humanoid agent. The mug is placed on top of the table, and the humanoid agent is behind the table from the camera’s perspective. We made several deliberate design choices to support our research goals. The minimalist, uncluttered environment reduces visual noise, ensuring the model learns to reason about the geometric relationships between the camera, agent, and object, rather than relying on spurious background correlations. The tabletop setting represents a common shared workspace for collaborative robotics. For the target object, we selected the SM_Mug_D1.usd model from the Isaac Sim props library. A mug is an ideal candidate as it is a common object that presents a clear interaction affordance (grasping), while its asymmetrical handle provides an unambiguous feature for learning precise 3D orientation. For the agent, we used the X-Bot.usd character from Mixamo, a stylized, mannequin-like form chosen for this proof-of-concept dataset. Finally, the use of a physically-based rendering engine provides realistic lighting, shadows, and reflections, which is an important step for grounding the task in a plausible physical context and facilitating future sim-to-real transfer.

2.1.2 Dataset Generation Pipeline

Building upon this base scene, our generation pipeline uses the NVIDIA Replicator engine to execute a Python script that implements domain randomisation. To ensure the model learns a robust representation, we programmatically randomised key pose parameters of the agent and object (specifically their planar translation and yaw) alongside key lighting parameters (see Table 1). This entire process was executed within the NVIDIA Synthetic Data Generation container [45] to ensure reproducibility. Within this container, Replicator’s programmable Randomizer modules performed the randomisation, while its Annotator modules extracted the precise 6-DOF ground-truth pose data required for training. This containerised

Table 1: Domain randomisation parameters for procedural scene generation. A uniform distribution was used to sample values for object pose and environmental lighting to create a visually diverse dataset and promote robust model training.

Category	Parameter	Target	Range
Pose	Translation (m)	Mug	$x \in [-0.3, 0.3], y \in [0, 0.3], z = 0.76$
	Rotation (Yaw, $^\circ$)	Mug	$[-180, 180]$
	Translation (m)	Humanoid	$x \in [-0.3, 0.3], y \in [1, 2], z = 0.0$
	Rotation (Yaw, $^\circ$)	Humanoid	$[-180, 180]$
Lighting	Intensity (arb. units)	Primary (Rect)	$[10000, 30000]$
	Colour Temp. (K)	Primary (Rect)	$[3500, 7500]$
	Position (m)	Primary (Rect)	$x, y \in [-5, 5], z \in [1.5, 3.0]$
	Rotation (Pitch, Roll, $^\circ$)	Primary (Rect)	Pitch $\in [-80, -10]$, Roll $\in [-180, 180]$
	Intensity (arb. units)	Ambient (Dome)	$[500, 1500]$
	Colour Temp. (K)	Ambient (Dome)	$[4000, 8000]$
	Rotation (Yaw, $^\circ$)	Ambient (Dome)	$[0, 360]$

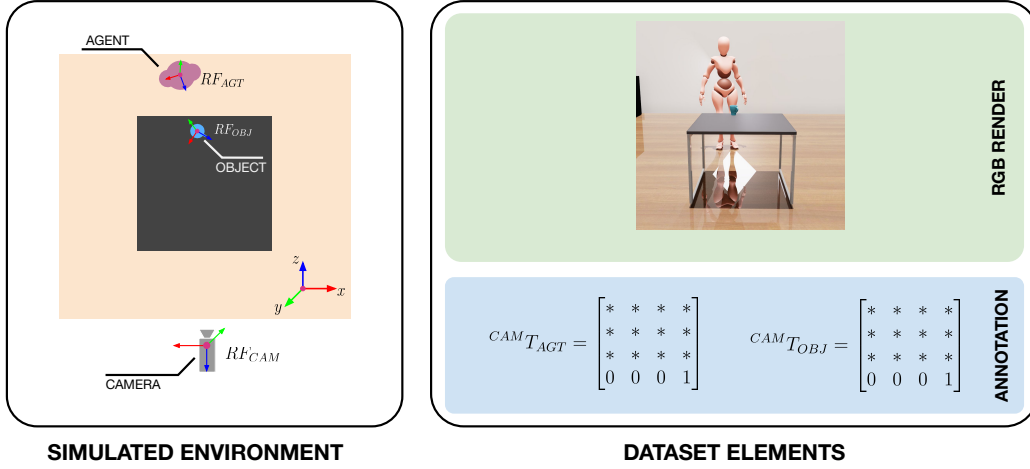


Figure 2: Overview of our data generation pipeline for the SynthVPT dataset. A scene (left) is procedurally generated in NVIDIA Replicator with precisely defined reference frames for the camera, agent, and object. The pipeline then outputs paired samples (right): a high fidelity RGB image and its ground-truth annotation, which contains the 6-DOF poses of the agent (${}^{CAM}T_{AGT}$) and object (${}^{CAM}T_{OBJ}$) relative to the camera.

workflow was operated in headless mode, enabling efficient data generation on a laptop equipped with an NVIDIA RTX 500 Ada Generation GPU.

For each of the 10,000 generated scenes, the Replicator captures a rendered RGB image (512×512 px) and the corresponding ground-truth transformation matrices for the mug, humanoid, and camera relative to the world frame. This raw output is then processed and structured into the final training dataset as described in Section 2.1.3.

2.1.3 Data Preprocessing

To transform the raw output from the NVIDIA Replicator pipeline into a structured format suitable for training, we developed a multi-step preprocessing script. The primary goal of this script is to parse the generated data, compute the precise 6-DOF pose of each object relative to the camera’s viewpoint, and structure the results into a dataset ready for supervised learning. The process is composed of three main stages: transformation calculation, pose decomposition, and dataset structuring.

Transformation Calculation: For each rendered frame, the pipeline outputs the camera’s pose in world coordinates, ${}^WT_{CAM}$, and the world pose of each entity of interest (i.e., the object and the agent), denoted generically as ${}^WT_{ENTITY}$. Our objective is to determine each entity’s pose from the camera’s perspective, ${}^{CAM}T_{ENTITY}$.

This is achieved through transformation composition. First, we compute the view matrix (${}^{CAM}T_W$), which transforms coordinates from the world frame to the camera frame by taking the inverse of the camera’s world pose matrix, as shown in Equation 1:

$${}^{CAM}T_W = ({}^WT_{CAM})^{-1} \quad (1)$$

Next, we compose this view matrix with an entity’s world pose to obtain its final transformation in the camera’s coordinate frame, as shown in Equation 2:

$${}^{CAM}T_{ENTITY} = {}^{CAM}T_W \cdot {}^WT_{ENTITY} \quad (2)$$

This process is applied to both the object (‘ENTITY = OBJ’) and the agent (‘ENTITY = AGT’), yielding the resulting matrices ${}^{CAM}T_{OBJ}$ and ${}^{CAM}T_{AGT}$. Together, these encapsulate the full 6-DOF ground-truth poses that serve as the training labels for each scene.

Pose Decomposition: The 4×4 transformation matrices, generically denoted as ${}^{CAM}T_{ENTITY}$, are decomposed into their translational and rotational components to create a more effective target for the model’s regression head. Directly regressing the 16 matrix values is ill-suited, as a standard loss function would fail to enforce the inherent mathematical constraints of a valid pose, such as the orthonormality of the rotation submatrix.

Our approach instead formulates the task as a multi-part prediction. The 3D translation vector would be regressed directly, with a loss defined by the Euclidean distance. The 3×3 rotation matrix is converted to a unit quaternion, which provides a continuous, non-singular representation suitable for stable training. This decomposition enables a composite loss function (a weighted sum of the translation and rotation errors) allowing for explicit control over the relative importance of positional and orientational accuracy.

The translation vector, $\vec{t} = [T_x, T_y, T_z]^T$, is extracted directly from the first three elements of the fourth column of ${}^{CAM}T_{ENTITY}$. These values are scaled by the scene’s `metersPerSceneUnit` parameter to ensure they are in metric units.

The rotation is represented by the upper-left 3×3 submatrix, R . To ensure a pure rotation matrix and handle any potential scaling artifacts from the simulation, we normalise the column vectors of R . This pure rotation matrix is then converted into a unit quaternion, $\vec{q} = [q_x, q_y, q_z, q_w]$. Quaternions are to be used as the regression target for orientation as they provide a continuous and non-singular representation, which is more stable for training deep learning models compared to Euler angles. This decomposition is performed for

both the object ($^{CAM}T_{OBJ}$) and agent ($^{CAM}T_{AGT}$) matrices to produce their respective 7-dimensional ground-truth vectors.

Dataset Structuring and Validation: A key step in this process is the creation of distinct training instances for each object of interest. For each of the 10,000 unique rendered images, two separate entries are created in the final dataset: one pairing the image with the humanoid’s pose and semantic label, and another pairing the same image with the mug’s pose and label.

This results in a final dataset of 20,000 instances, which is then partitioned into training (16,000), validation (2,000), and testing (2,000) sets. Critically, this split is performed based on the 10,000 unique images, not the 20,000 individual instances. This strategy prevents data leakage by ensuring that an image seen during training will never appear in the validation or test sets, even with a different object label.

As a final integrity check, we compute an SHA256 hash of the pixel data for every image to programmatically verify that no identical images exist across these partitions. The final, structured dataset is then saved as a Hugging Face DatasetDict, providing a standardised and efficient format for our training pipeline.

2.2 Spatial Grounding Module

The core of our proposed framework is the Spatial Grounding Module (SGM), a component designed to provide the foundational spatial reasoning that underpins embodied social cognition (see Figure 3). While the ultimate goal of VPT is social (understanding another’s viewpoint) this ability must first be grounded in a precise metric representation of the physical world.

Conventional 6-DOF pose estimators (e.g., LINEMOD [27] and PoseCNN [57]) rely on extensive object-specific datasets, CAD models, and rigid architectures, which limit their generalisation to new objects or scenarios. In contrast, we propose that a pre-trained VLM can be lightly fine-tuned to acquire metric spatial grounding using only a small synthetic dataset. While the model itself remains conceptual, this approach is theoretically well-motivated: multimodal pre-training in VLMs encodes rich semantic and relational structure [3, 47], which can be adapted to predict geometric quantities such as translation and rotation. Therefore, even minimal fine-tuning may suffice to link linguistic and visual representations within a physically consistent spatial frame: an essential prerequisite for VPT and more broadly, embodied social cognition.

To bridge this gap, we propose an SGM built by adapting a pre-trained VLM for the specific task of 6-DOF pose regression. Our approach would utilise Parameter-Efficient Fine-Tuning (PEFT) [25] and a specialised Multi-Layer Perception (MLP) [55] regression head to map the model’s rich visual features to a 7-dimensional pose vector. This allows the SGM to produce the precise geometric data (the camera-relative poses of both the agent and the object) which serves as the essential input for any downstream social reasoning about another’s perspective.

2.2.1 Foundation Model

The core of our proposed SGM architecture is the Smo1VLM-500M-Instruct model, an open-source instance of the Idefics3 [38] family. This compact 500 million parameter model was chosen for its suitability for our target application in embodied robotics, where

its small size is expected to enable low-latency inference on power-constrained hardware [42]. The model consists of a SigLIP [59] vision encoder, a language model, and a connector module that projects visual features into the language model’s embedding space.

2.2.2 Architecture for 6-DOF Pose Estimation

To adapt the VLM for regression, we will prompt the model with both an image and a task-specific text query (e.g., “What is the 6-DOF pose of the mug?”). The process for deriving a 7D pose vector from an input image I and prompt P is as follows:

- (1) **VLM Processing:** The input image I and text prompt P are passed through the full VLM, which generates a sequence of hidden states from its final language model layer.
- (2) **Feature Extraction:** We will extract the hidden state corresponding to the final token of the input sequence, $h_{last} \in \mathbb{R}^{D_{hidden}}$, where $D_{hidden} = 960$. This vector serves as a holistic representation of the image and text prompt.
- (3) **Regression Head:** The feature vector h_{last} is then passed through an MLP which serves as the regression head. The head consists of a Layer Normalisation step followed by a series of fully connected layers with ReLU activations. The MLP architecture is defined by the following layer transitions: $960 \rightarrow 1024 \rightarrow 512 \rightarrow 128 \rightarrow 7$.
- (4) **Output:** The final layer outputs a 7-dimensional vector $\hat{\xi} = [\hat{T}_x, \hat{T}_y, \hat{T}_z, \hat{q}_x, \hat{q}_y, \hat{q}_z, \hat{q}_w]$, which corresponds to the predicted translation coordinates and rotation quaternion.

2.2.3 Training Objective

The model is trained to minimise a composite loss function, \mathcal{L}_{total} , which is a weighted sum of a translation loss and a rotation loss (denoted by \mathcal{L}_{trans} and \mathcal{L}_{rot} respectively). This allows us to explicitly balance the learning of positional and orientational accuracy, as defined in Equation 3.

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{trans} + \beta \mathcal{L}_{rot} \quad (3)$$

The hyperparameters α and β will be determined empirically during training to balance the two loss components, as their relative scales can influence gradient updates and convergence.

Translation Loss: For the translation component, we use the Mean Euclidean Distance (L2 norm). This loss function is a direct and intuitive measure of the physical error in 3D space. It provides a smooth gradient for the optimiser and penalises large positional errors more heavily, encouraging the model to achieve high positional accuracy. The loss for a batch of B samples is given Equation 4, where $\hat{\mathbf{t}}_i$ is the predicted translation vector and \mathbf{t}_i is the corresponding ground-truth translation vector for the i -th sample.

$$\mathcal{L}_{trans} = \frac{1}{B} \sum_{i=1}^B \|\hat{\mathbf{t}}_i - \mathbf{t}_i\|_2 \quad (4)$$

Rotation Loss: For rotation, we adopt a loss based on unit quaternions to avoid the inherent problems of other 3D rotation representations, such as Euler angles. Euler angles suffer from discontinuities and the critical issue of gimbal lock, which can create unstable gradients and hinder model training. Quaternions provide a continuous and non-singular representation of orientation, making them a much more stable target for a neural network to learn. Our rotation loss,

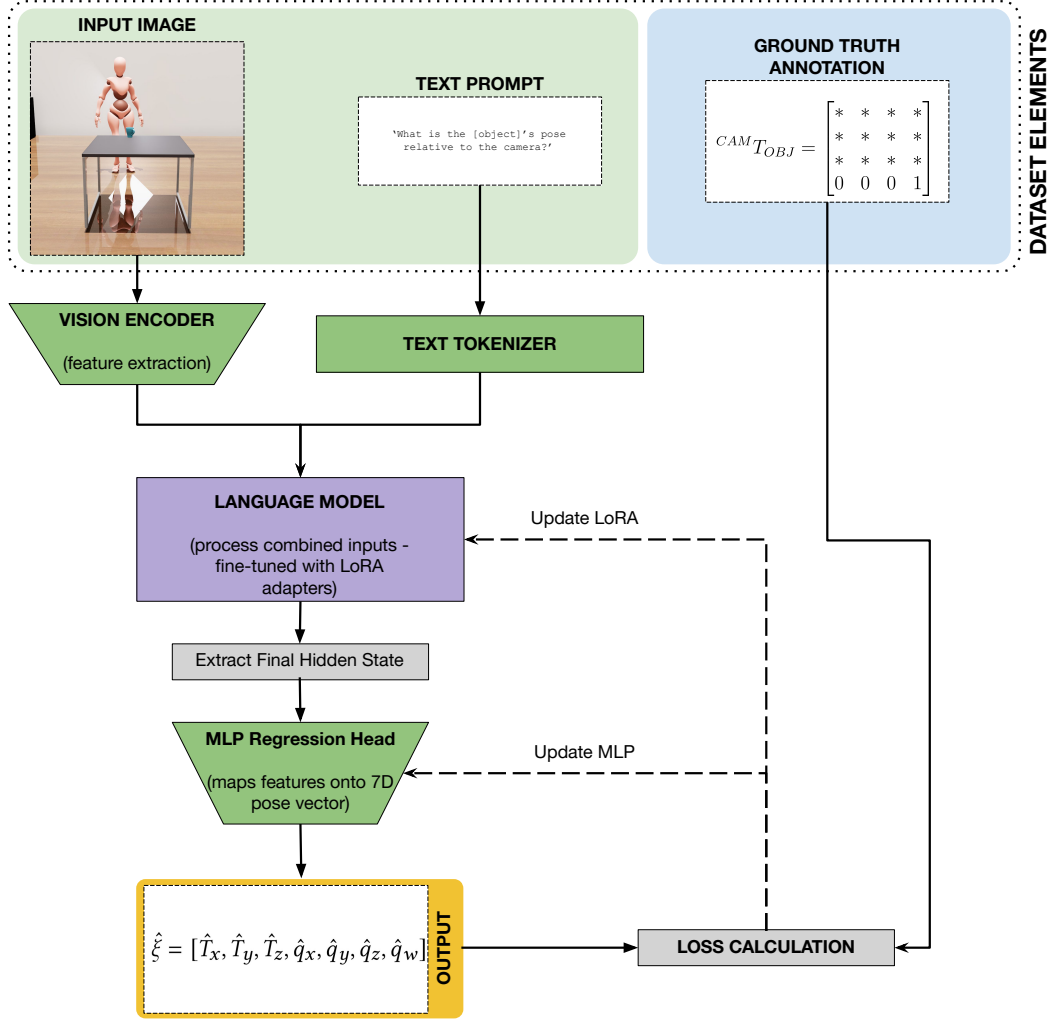


Figure 3: The proposed architecture for the SGM, the foundational component of our VPT framework. This module is responsible for grounding the VLMs understanding, in metric space. It takes an RGB image and a text prompt as input and regresses a camera-relative 7D pose vector ($\hat{\xi}$). During training, a composite loss between the prediction and the ground-truth annotation updates the MLP head and LoRA adapters. The pose predicted by this module serves as the necessary input for downstream perspective mapping tasks.

\mathcal{L}_{rot} , is defined as the L2 norm between the predicted and ground-truth unit quaternions. To account for the double-cover property of quaternions (where both \mathbf{q} and its negative $-\mathbf{q}$ represent the same rotation) we minimise the shorter of the two possible distances. This ensures the loss is unambiguous and reflects the true angular error, as seen in Equation 5, where $\hat{\mathbf{q}}_i$ and \mathbf{q}_i are the predicted and ground-truth quaternions, respectively.

$$\mathcal{L}_{rot} = \frac{1}{B} \sum_{i=1}^B \min(\|\hat{\mathbf{q}}_i - \mathbf{q}_i\|_2, \|\hat{\mathbf{q}}_i + \mathbf{q}_i\|_2) \quad (5)$$

2.2.4 Implementation and Training Details

We propose implementing the model using PyTorch [32] and the HuggingFace Transformers library [33]. To make fine-tuning computationally tractable, several efficiency techniques can be employed. The base VLM’s weights will be loaded using 4-bit nf4 quantization. We will then use Low-Rank Adaptation (LoRA) [30], a PEFT method, to adapt the model. LoRA adapters with a rank of 8 will be inserted into the attention projection layers of the VLM.

During training, only the LoRA adapter weights and the randomly initialised MLP regression head will be updated. This strategy drastically reduces the number of trainable parameters, enabling us to fine-tune the large model on modest hardware. We will use the AdamW

optimiser with a learning rate determined empirically through hyperparameter tuning, and employ bfloat16 mixed-precision training to further enhance computational efficiency.

2.3 Perspective Mapping

The final stage of our proposed framework is designed to integrate the outputs from the SGM to perform the final VPT computation. We deliberately propose structuring this stage as a deterministic geometric calculation rather than an end-to-end learned function. This modular, decomposable approach offers several key advantages.

- i. **Interpretability:** If the final perspective inference were incorrect, one could isolate the source of the error to either the object pose estimation, the agent pose estimation, or an error in the implementation of the transformation logic itself.
- ii. **Data Efficiency:** Structuring the framework with distinct Spatial Grounding and Perspective Mapping modules is a deliberate design choice, essential for our intended application domain, robotics. This modularity is particularly critical for the compact models demanded by real-world robotics applications requiring low-latency inference. An end-to-end system would force such a model to learn a single, highly complex function mapping pixels directly to the final agent-relative pose (${}^{AGT}\hat{T}_{OBJ}$). This function implicitly combines two fundamentally different problems: the ambiguous perceptual challenge of inferring 3D structure from a 2D image, and the rigid, mathematical rules of geometric perspective transformation. Forcing a single network to approximate this entire composite function from data is inefficient, as it must expend its precious and limited learning capacity discovering a relationship that is already perfectly defined by geometry. Our modular approach decouples these problems. We propose tasking the model with a more constrained and stable learning target: the direct perceptual grounding of objects within a single reference frame. By then applying the geometric transformation as a deterministic final step, we inject perfect domain knowledge into the system. This offloads the most complex relational logic, allowing the model to dedicate its entire capacity to the perceptual task it is best suited for. This is a key strategy for enabling smaller models to successfully perform complex spatial reasoning on resource-constrained hardware.
- iii. **Reusability:** By decoupling spatial perception from relational reasoning, the SGM is not intrinsically tied to the VPT task. Its camera-relative pose estimations (${}^{CAM}\hat{T}_{OBJ}$ and ${}^{CAM}\hat{T}_{AGT}$) serve as foundational spatial outputs, making them directly applicable to other critical robotics tasks such as grasp planning and social navigation, independent of the final perspective transformation.
- iv. **Scalability:** The modular design is inherently scalable to complex, multi-agent and multi-object scenarios. The SGM functions as a universal pose estimator that can be queried independently for any number of entities in a scene. The resulting camera-relative poses can then be combined by the deterministic mapping stage to compute any desired pairwise perspective (e.g., agent A’s view of object B, agent C’s view of object D) without requiring the model to be retrained or

to learn a combinatorial number of relational mappings. An end-to-end model would lack this compositional flexibility.

The proposed mapping process itself would be a transformation composition that requires no learnable parameters. It would take the two poses predicted by the SGM as input: the pose of the object relative to the camera, ${}^{CAM}\hat{T}_{OBJ}$, and the pose of the agent relative to the camera, ${}^{CAM}\hat{T}_{AGT}$.

To find the object’s pose from the agent’s perspective, the frame of reference must be changed from the camera to the agent. The first step is to compute the transformation from the agent’s frame to the camera’s frame, ${}^{AGT}\hat{T}_{CAM}$, by taking the inverse of the predicted agent pose:

$${}^{AGT}\hat{T}_{CAM} = ({}^{CAM}\hat{T}_{AGT})^{-1} \quad (6)$$

With this transformation, the final object pose relative to the agent, ${}^{AGT}\hat{T}_{OBJ}$, can be found by pre-multiplying the object’s camera-relative pose by ${}^{AGT}\hat{T}_{CAM}$:

$${}^{AGT}\hat{T}_{OBJ} = {}^{AGT}\hat{T}_{CAM} \cdot {}^{CAM}\hat{T}_{OBJ} \quad (7)$$

The resulting 4×4 matrix, ${}^{AGT}\hat{T}_{OBJ}$, would provide a complete, metric description of the object’s position and orientation from the agent’s viewpoint. By decoupling the complex perceptual challenge from the solved problem of geometric transformation, our framework is designed to produce a robust and verifiable output. This final transformation provides the necessary geometric grounding for perspective aware downstream tasks, such as interpreting perspective-dependent instructions and reasoning about what human partners can see, ultimately enabling safer and more intuitive collaboration.

3 Discussion

In this work, we present a framework and a proof-of-concept dataset, SynthVPT, for teaching VLMs to perform VPT. We argue that for VLMs to serve as the socio-cognitive enabler of embodied agents in HRI, their powerful semantic understanding must be grounded in the geometric reality of the physical world. By reformulating VPT as a 6-DOF pose regression task, we aim to shift the capabilities of these models beyond categorical spatial judgments (e.g., "left of") towards the precise, metric reasoning essential for embodied tasks like resolving ambiguous commands and understanding occlusions.

A key implication of our work lies in the proposed modular architecture, which deliberately decouples the perceptual challenge of 3D understanding from the solved problem of geometric transformation. An end-to-end model would be forced to learn both perception and projective geometry from scratch, an inefficient use of its limited capacity. Our approach instead tasks the VLM with a more focused and stable learning target: predicting the camera-relative pose of objects in a scene. The deterministic *Perspective Mapping* stage then injects perfect, verifiable domain knowledge into the system. This approach not only improves data efficiency and interpretability but also creates a more reusable and scalable system. The SGM, once trained, acts as a general-purpose pose estimator that can be repurposed for other robotics tasks, such as grasp planning, without being intrinsically tied to VPT.

Grounding VPT in metric space has profound implications for HRI. It is the foundational skill that would allow a robot to move beyond literal interpretations of language and towards a genuine

shared understanding. For instance, it could resolve perspective-dependent instructions ("hand me the tool on *my* right"), reason about a human's visual field to understand occlusions ("I know you can't see the warning light from your angle"), and ultimately engage in safer and more fluid physical collaboration. Without this metric grounding, any higher-level social reasoning about a human's beliefs, desires, or intentions would remain ungrounded from the physical context of the interaction.

This work therefore provides a tangible first step towards developing the *Social Foundation Models* we propose. While true social intelligence requires more than just geometric understanding, we argue that metric spatial reasoning is a necessary, albeit not sufficient, prerequisite. It establishes the stable, physical foundation upon which more complex social behaviours can be composed from a library of learned socio-cognitive primitives. By demonstrating a scalable, data-driven blueprint for teaching this foundational skill, we lay the groundwork for a new class of models capable of perceiving and acting within a shared, human-centric world.

3.1 Limitations and Scope

We acknowledge several key limitations that define the scope of this initial work. These limitations not only frame our current contributions but also directly inform our future research.

3.1.1 Synthetic Data and Generalisation

The most significant challenge is the sim-to-real gap. While our physically-based rendering is a step towards realism, the proposed model would be exclusively trained on synthetic data. Its ability to generalise to the noise, clutter, and visual complexity of real-world imagery will need to be empirically validated. This challenge is compounded by the scope of our initial SynthVPT dataset. As a proof-of-concept, its deliberate simplicity (a single scene archetype, object class, and stylised avatar) constrains its immediate generalisability. Therefore, a key priority for future work is to systematically extend the dataset to overcome these challenges. This will involve incorporating a diverse range of photorealistic assets, including cluttered environments representative of our target domains (e.g. social robotics), a large library of everyday objects, and varied human models that account for differences in appearance and attire.

3.1.2 Task Formulation and Interaction Dynamics

Our current formulation models a constrained interaction: a single agent's perspective on a single object from a static camera view. This simplification omits several key aspects of real-world HRI:

Static Scenes and Temporal Dynamics: The use of static images means our framework would reason about a single moment in time, omitting the temporal dynamics that are essential for advanced social cognition. Core collaborative abilities, such as anticipating a human's intentions or engaging in fluid joint action, are fundamentally temporal and rely on understanding motion, trajectories, and how perspectives evolve. However, our SGM proposal is designed as a foundational component to enable this future work. Rather than being a limitation, this modularity is a strength. The 7D pose vectors our SGM generates could serve as a powerful, low-dimensional state representation at each time step (t). This sequence of state

vectors could then become the ideal input for a downstream temporal model. For instance, a secondary Transformer [11] or Long Short-Term Memory (LSTM) [28] focused on trajectory forecasting or human intent prediction. In this architecture, our SGM would function as a robust perception module offloading the difficult per-frame spatial grounding task and allowing the temporal model to focus purely on learning the dynamics of the interaction.

Prompt Dependency and Agent-Tool Interaction: A key aspect of our framework is that the SGM is conceived not as a standalone system, but as a specialised tool to be used by a higher-level AI agent [46, 49]. In its current formulation, this tool is queried via an explicit text prompt (e.g., "What is the 6-DOF pose of the mug?"), a simplification that is a significant constraint for fluid HRI. Real-world collaboration is not a turn-based, command-driven process; it relies on implicit social cues to direct attention and action [5, 13].

A key direction for future work would be to empower the agent to query this tool autonomously, using an implicit attentional trigger. A promising approach, grounded in our group's research is to envision a gaze estimation module serving as a perception front-end for the agent. Such a module could identify the human's object of attention in the shared workspace, allowing the agent to internally generate the appropriate query for its SGM tool. This would shift the interaction from a rigid, command-based paradigm to a more natural, attention-driven one. This is not merely a technical solution for triggering inference; it is a fundamental requirement for effective social interaction. Research has demonstrated that the very act of a robot establishing eye contact has a positive impact on how it is perceived by a human partner [1]. It increases feelings of engagement and attributions of human-likeness, even when the robot's gaze is not predictive of a subsequent task [36]. This highlights the importance of grounding social reasoning not just in the geometric accuracy our framework proposes, but also in the reciprocal social signals that define natural interaction.

Single Agent Focus: The current proof-of-concept dataset is limited to a single agent's perspective, which simplifies the problem to a pairwise relationship between one agent and one object. True collaborative intelligence, however, requires reasoning within a multi-agent context. This would involve not just understanding multiple viewpoints, but also higher-order social reasoning about shared vs. private knowledge (e.g., "Agent A knows that Agent B cannot see the object"). While our current dataset does not facilitate this, the modularity of our proposed framework is a key architectural choice designed for this future scalability. The SGM provides the foundational geometric data, and the deterministic mapping stage offers the compositional logic needed to begin addressing these more complex multi-agent dynamics.

3.1.3 Model Interpretability and Scale

Model Scale and Deployment: Our proposal to use a compact 500M parameter VLM is a pragmatic choice, balancing the high performance of larger models with the practical constraints of robotic deployment. While cloud-based inference offers a solution to on-board computational limits, our focus on a compact, locally-deployable model is motivated by the demands of safe, robust, and ethical HRI. Many collaborative tasks require real-time reactivity, such as stopping before a collision, or responding to a sudden human gesture, that cannot tolerate the network latency inherent in cloud inference.

Local inference also ensures operational autonomy, guaranteeing functionality without a stable network connection. Crucially, it mitigates the significant privacy risks [31] associated with continuously transmitting potentially sensitive visual data of human collaborators to third-party servers.

Interpretability and Safety: More fundamentally, regardless of scale or deployment location, the issue of interpretability remains a challenge. While our proposed model could learn the function to predict a pose with high accuracy, its internal reasoning would remain a black box. We are proposing to teach it to map pixels to coordinates, a form of functional approximation that is not necessarily equivalent to forming a human-like, interpretable geometric understanding of the scene. Developing methods to scrutinise and validate the model’s internal geometric “understanding” will be crucial for ensuring its reliability and safety in critical HRI applications [4].

3.2 Future Work

The immediate priority is to execute the proposed fine-tuning to establish a performance baseline for the SGM on our synthetic dataset. This initial validation will involve a rigorous benchmarking process against two key baselines: (i) a conventional CNN-based pose regressor (e.g., a ResNet backbone [26]) trained on the same data, and (ii) an ablation study of our model trained from scratch without pre-trained weights. Following this, we will evaluate the model’s zero-shot generalisation capabilities on a small test set of novel objects to test the key hypothesised advantage of our VLM-based approach.

Success in these initial evaluations will motivate the most critical long-term direction: sim-to-real validation and deployment on a physical robotic platform. This involves quantifying the model’s performance in real-world conditions with sensor noise and visual clutter. This step is the true test of our approach and will guide the systematic expansion of the SynthVPT dataset with more diverse assets and contexts to close the reality gap.

Once validated in the real world, the final research direction will be to demonstrate the SGM’s downstream utility in tasks representative of our target application domains, thereby taking the first concrete steps towards realising the *Social Foundation Models* we propose. For social robotics, this includes using the predicted poses to resolve perspective-dependent commands for grasp planning [39]. For autonomous navigation, this involves inferring a pedestrian’s visual field from their pose to enable more robust intent prediction [60]. Success in these embodied tasks would provide strong evidence that our methodology—grounding VLMs in precise geometric data—is a foundational and scalable pathway towards developing more sophisticated and socially aware robotic agents.

4 Conclusion

We argue that for embodied agents to become truly collaborative partners, they require, in addition to the powerful existing semantic capabilities present in state of the art multimodal models, a deep understanding of the physical and social context of an interaction. To this end, we propose *Social Foundation Models*: a new class of models designed to master crucial socio-cognitive skills. The core of these social primitives is VPT, the ability to represent the world

from another’s physical viewpoint. However, we argue that for VPT to be functional in HRI, it demands a metric precision that current VLMs inherently lack due to their training on web-scale data.

Our work provides a methodology for bridging this critical data gap. We contribute SynthVPT, a novel dataset of geometrically precise RGB images, and proposed a model architecture that reformulates VPT as a 6-DOF pose regression task. This approach deliberately decouples the perceptual challenge of 3D understanding from the deterministic logic of geometric transformation, designed as a data-efficient, interpretable, and reusable system. By tasking a VLM with predicting camera-relative poses, we establish a foundational general-purpose spatial reasoner.

By grounding powerful semantic models in precise geometric data, this methodology is therefore a tangible and necessary first step towards realising the *Social Foundation Models* we propose. It provides the stable physical foundation upon which more abstract socio-cognitive skills can be built, leading to a new generation of robots that move beyond simple instruction-following and towards a genuine, shared reality with their human partners.

Dataset Availability

To support further research, we release the SynthVPT dataset [16]. <https://huggingface.co/datasets/jwgcurrie/SynthVPT>.

Acknowledgments

We are very grateful to Prof. Patric Bach, Dr Maria Elena Giannaccini, and our colleagues at the S4HRI Unit (IIT) and the Action Prediction Lab (University of Aberdeen) for their valuable and stimulating discussions regarding this work. Finally, we thank Mr Francesco Gervino for introducing us to the concept and history of *Flatland*.

This work has received support from the Project “*Future Artificial Intelligence Research (hereafter FAIR)*”, PE000013 funded by the European Union - NextGenerationEU PNRR MUR - M4C2 - Investimento 1.3 - Avviso Creazione di “*Partenariati estesi alle università, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base*” CUP J53C22003010006.

References

- [1] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *J. Hum.-Robot Interact.* 6, 1 (May 2017), 25–63. <https://doi.org/10.5898/JHRI.6.1.Admoni>
- [2] Naveed Ahmed, Imad Afyouni, Hamzah Dabool, and Zaher Al Aghbari. 2024. A systemic survey of the Omniverse platform and its applications in data generation, simulation and metaverse. *Frontiers in Computer Science* 6 (2024), 1423129.
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [4] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, Montreal, Canada, May 13–17, 2019. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.
- [5] Patric Bach and Kimberley C Schenke. 2017. Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass* 11, 7 (2017), e12312.
- [6] C Daniel Batson, Shannon Early, and Giovanni Salvareni. 1997. Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and social psychology bulletin* 23, 7 (1997), 751–758.
- [7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma

- Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Dombouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kavin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] <https://arxiv.org/abs/2108.07258>
- [8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishk Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818 [cs.RO] <https://arxiv.org/abs/2307.15818>
- [9] Nicholas Budny, Kia Ghods, Declan Campbell, Raja Marjeh, Amogh Joshi, Sreejan Kumar, Jonathan D. Cohen, Taylor W. Webb, and Thomas L. Griffiths. 2025. Visual serial processing deficits explain divergences in human and VLM reasoning. arXiv:2509.25142 [cs.AI] <https://arxiv.org/abs/2509.25142>
- [10] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14455–14465.
- [11] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. arXiv:2106.01345 [cs.LG] <https://arxiv.org/abs/2106.01345>
- [12] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025. Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas. arXiv:2503.01773 [cs.CL] <https://arxiv.org/abs/2503.01773>
- [13] Joel Currie, Maria Elena Giannaccini, and Patric Bach. 2024. Sonic Sleight of Hand: Sound induces illusory distortions in the perception and prediction of robot action. *International Journal of Social Robotics* (2024), 1–19.
- [14] Joel Currie, Katrina Louise McDonough, Agnieszka Wykowska, Maria Elena Giannaccini, and Patric Bach. 2024. Mind Meld or Mismatch: A Comparison of Visual Perspective Taking Towards Humans and Robots in Face-to-Face Interactions. <https://doi.org/10.31219/osf.io/zh7sg>
- [15] Joel Currie, Katrina Louise McDonough, Agnieszka Wykowska, Maria Elena Giannaccini, and Patric Bach. 2024. More Than Meets the Eye? An Experimental Design to Test Robot Visual Perspective-Taking Facilitators Beyond Mere-Appearance. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 359–363. <https://doi.org/10.1145/3610978.3640684>
- [16] Joel Currie, Gioele Migno, Enrico Piacenti, Maria Elena Giannaccini, Patric Bach, Davide De Tommaso, and Agnieszka Wykowska. 2025. synthetic-distance (Revision c86eff8). <https://doi.org/10.57967/hf/5351>
- [17] Joel Currie, Gioele Migno, Enrico Piacenti, Maria Elena Giannaccini, Patric Bach, Davide De Tommaso, and Agnieszka Wykowska. 2025. Towards Embodied Cognition in Robots via Spatially Grounded Synthetic Worlds. arXiv:2505.14366 [cs.AI] <https://arxiv.org/abs/2505.14366>
- [18] Fethiye Irmak Doğan, Sarah Gillet, Elizabeth J. Carter, and Iolanda Leite. 2020. The impact of adding perspective-taking to spatial referencing during human-robot interaction. *Robotics and Autonomous Systems* 134 (2020), 103654. <https://doi.org/10.1016/j.robot.2020.103654>
- [19] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiye Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. 2025. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research* 44, 5 (2025), 701–739. <https://doi.org/10.1177/02783649241281508> arXiv:https://doi.org/10.1177/02783649241281508
- [20] Tobias Fischer and Yiannis Demiris. 2016. Markerless perspective taking for humanoid robots in unconstrained environments. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 3309–3316. <https://doi.org/10.1109/ICRA.2016.7487504>
- [21] Martin Freundlieb, Ágnes M Kovács, and Natalie Sebanz. 2016. When do humans spontaneously adopt another’s visuospatial perspective? *Journal of experimental psychology: human perception and performance* 42, 3 (2016), 401.
- [22] Tiziano Furlanetto, Cristina Becchio, Dana Samson, and Ian Apperly. 2016. Alter-centric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance* 42, 2 (2016), 158.
- [23] Qingying Gao, Yijiang Li, Haiyun Lyu, Haoran Sun, Dezhi Luo, and Hokin Deng. [n.d.]. Vision Language Models See What You Want but not What You See. <https://doi.org/10.48550/arXiv.2410.00324> arXiv:2410.00324 [cs]
- [24] Gracjan Góral, Alicja Ziarko, Michał Nauman, and Maciej Wolczyk. [n.d.]. Seeing Through Their Eyes: Evaluating Visual Perspective Taking in Vision Language Models. <https://doi.org/10.48550/arXiv.2409.12969> arXiv:2409.12969 [cs]
- [25] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024).
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [27] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. 2012. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*. Springer, 548–562.
- [28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [29] Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2025. EgoSocialArena: Benchmarking the Social Intelligence of Large Language Models from a First-person Perspective. arXiv:2410.06195 [cs.CL] <https://arxiv.org/abs/2410.06195>
- [30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.
- [31] Xiangyu Hu and Zhenyu Xu. 2025. Large language and vision-language models for robot: safety challenges, mitigation strategies and future directions. *Industrial Robot: the international journal of robotics research and application* (2025).
- [32] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. 2021. PyTorch. In *Programming with TensorFlow: solution for edge computing applications*. Springer, 87–104.
- [33] Shashank Mohan Jain. 2022. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*. Springer, 51–67.
- [34] A. S. Johnson, B. Clarke, and C. Jones. 2015. Robotic Visual Perspective Taking via Geometric Reasoning. *IEEE Transactions on Robotics* 31, 6 (2015), 1352–1367. <https://doi.org/10.1109/RO.2015.2495016> ISSN: 1552-3098.
- [35] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. 2024. Real-world robot applications of foundation models: a review. *Advanced Robotics* 38, 18 (2024), 1232–1254. <https://doi.org/10.1080/01691864.2024.2408593> arXiv:https://doi.org/10.1080/01691864.2024.2408593
- [36] Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhanoff, Giorgio Metta, and Agnieszka Wykowska. 2021. It’s in the eyes: The engaging role of eye contact in HRI. *International Journal of Social Robotics* 13, 3 (2021), 525–535.
- [37] Maria Kozhevnikov, Michael A. Motes, Bjørn Rasch, and Olessia Blajenkova. 2006. Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. *Applied Cognitive Psychology* 20, 3 (2006), 397–417. <https://doi.org/10.1002/acp.1192> arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.1192
- [38] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. arXiv:2408.12637 [cs.CV] <https://arxiv.org/abs/2408.12637>
- [39] Haichao Liu, Sikai Guo, Pengfei Mai, Jiahang Cao, Haoang Li, and Jun Ma. 2025. RoboDexVLM: Visual Language Model-Enabled Task Planning and Motion Control for Dexterous Robot Manipulation. arXiv:2503.01616 [cs.RO] <https://arxiv.org/abs/2503.01616>
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914f369fe6de0-Paper-Conference.pdf
- [41] Dezhi Luo, Yijiang Li, and Hokin Deng. 2025. The Philosophical Foundations of Growing AI Like A Child. *arXiv preprint arXiv:2502.10742* (2025).

- [42] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. SmolVLM: Redefining small and efficient multimodal models. *arXiv:2504.05299* [cs.AI] <https://arxiv.org/abs/2504.05299>
- [43] Luis Felipe Marin-Urias, E Akin Sisbot, and Rachid Alami. 2008. Geometric tools for perspective taking for human-robot interaction. In *2008 Seventh Mexican International Conference on Artificial Intelligence*. IEEE, 243–249.
- [44] Bradley D Mattan, Pia Rotshtein, and Kimberly A Quinn. 2016. Empathy and visual perspective-taking performance. *Cognitive neuroscience* 7, 1-4 (2016), 170–181.
- [45] NVIDIA. 2025. NVIDIA Omniverse Synthetic Data Generation Container. <https://catalog.ngc.nvidia.com/orgs/nvidia/containers/ov-synthetic-data-generation>. Accessed: 17-10-2025.
- [46] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. 2024. Tool learning with foundation models. *Comput. Surveys* 57, 4 (2024), 1–40.
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*. 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html> ICML 2021.
- [48] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. 2024. SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models. *arXiv preprint arXiv:2412.07755* (2024).
- [49] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv:2302.04761* [cs.CL] <https://arxiv.org/abs/2302.04761>
- [50] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. 2023. ViNT: A Foundation Model for Visual Navigation. *arXiv:2306.14846* [cs.RO] <https://arxiv.org/abs/2306.14846>
- [51] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. *arXiv:2305.14763* [cs.CL] <https://arxiv.org/abs/2305.14763>
- [52] Ritvik Singh, Jingzhou Liu, Karl Van Wyk, Yu-Wei Chao, Jean-Francois Lafleche, Florian Shkurti, Nathan Ratliff, and Ankur Handa. 2024. Synthetica: Large Scale Synthetic Data for Robot Perception. *arXiv:2410.21153* [cs.CV] <https://arxiv.org/abs/2410.21153>
- [53] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2024. RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics. *arXiv preprint arXiv:2411.16537* (2024).
- [54] Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, and Yanyi Chen. 2025. A Cognitive Evaluation Benchmark of Image Reasoning and Description for Large Vision-Language Models. *arXiv:2402.18409* [cs.AI] <https://arxiv.org/abs/2402.18409>
- [55] Hind Taud and Jean-Francois Mas. 2017. Multilayer perceptron (MLP). In *Geomatic approaches for modeling land change scenarios*. Springer, 451–455.
- [56] J.G. Trafton, N.L. Cassimatis, M.D. Bugajska, D.P. Brock, F.E. Mintz, and A.C. Schultz. 2005. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 35, 4 (July 2005), 460–470. <https://doi.org/10.1109/TSMCA.2005.850592> Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.
- [57] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2018. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *arXiv:1711.00199* [cs.CV] <https://arxiv.org/abs/1711.00199>
- [58] Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. 2024. A Survey on Robotics with Foundation Models: toward Embodied AI. *arXiv:2402.02385* [cs.RO] <https://arxiv.org/abs/2402.02385>
- [59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sig-moid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.
- [60] Shucheng Zhang, Yan Shi, Bingzhang Wang, Yuang Zhang, Muhammad Mon-jurul Karim, Kehua Chen, Chenxi Liu, Mehrdad Nasri, and Yinhai Wang. 2025. A Comprehensive Review on Artificial Intelligence Empowered Solutions for Enhancing Pedestrian and Cyclist Safety. *arXiv:2510.03314* [cs.CV] <https://arxiv.org/abs/2510.03314>