Disentangling Grasp-Object Representations in the Latent Space: Toward Brain-like Affordances for Machines

Siavash Mortaz Hejri[1][0000-1111-2222-3333], Hamed Pourfannan[1][0000-0002-5589-2961], Ruidong Ma[1][0000-0002-8035-5746], Alejandro Jimenez Rodriguez[1][0000-0001-7172-1794], and Alessandro Di Nuovo[1][0000-0003-2677-2650]

School of Computing, Sheffield Hallam University, Sheffield, United Kingdom

Abstract. Humans perceive the world through their bodies. The theory of object affordances suggests that when encountering an object, our brain encodes it not only based on its physical properties but also according to how we intend to use it. Decades of foundational research in neuroscience indicate that object properties are associated with distinct regions of the sensorimotor cortex, depending on the grasp type they tend to activate. In this study, we trained a Conditional Variational Autoencoder (CVAE) on the HO-3D_v3 dataset to reconstruct hand poses conditioned on object properties. Principal Component Analysis (PCA), clustering, and visualization of the model's latent space revealed structured patterns for the abstract representation of the hand, which were distinctly organized according to object associations. This bears a notable resemblance to neural strategies observed in the human sensorimotor cortex for representing object-grasp relationships. This finding supports the notion that artificial intelligence systems can develop brain-like latent representations of object affordances. Such representations could significantly enhance robotic control in the future by enabling real-time motor planning for high-degree-of-freedom humanoid hand actions in an abstract latent space, bypassing the need for low-level pixel- and joint-level computations.

Keywords: Conditional Variational Autoencoders (CVAEs) · Latent Space Analysis · Principle Component Analysis · Grasp Embeddings · Object Affordances

1 Introduction

The concept of Embodied Cognition implies that human cognition extends beyond the brain to the body and the environment of the agent, all of which contribute to how we perceive the world and act upon it [1]. Despite different arguments over the extent of this expansion, rich empirical data support the core claims of this theory[2, 3]. One idea, theoretically tied to the concept of embodied cognition, is the object affordances. Originally advocated by Gibson in 1979

[4], this theory suggests that environments and objects hold within themselves specific possibilities for action that they "afford". Based on this approach, each object is primed for specific types of action based on its functionality, which the animal learns through experience or observing others. There is solid neuroscientific evidence in support of the fact that objects are represented based on their perceived functionality and related motor plans [5]. For instance, it is shown that when observing an object, both visual and motor signals are generated in the brain, regardless of the intention to act upon that specific object [6]. Other behavioral studies suggest that when presented with an object, the motor plans appropriate for that object shape (precision versus power grips) are potentiated, violation of which, through creative task design, results in increased reaction time in the subject to readjust the plans [7].

At the neural level, affordance representation in the brain during grasping action has been examined through brain-imaging studies, which show that both object properties and grasp types can be decoded from EEG signals before movement execution, significantly above chance [8]. By systematically varying object features (e.g., shape, size, orientation) and grasp types (e.g., precision, power, parallel grips), researchers demonstrated that distinct nodes within the sensorimotor cortex encode different components of grasping—such as finger configuration, object shape, and grip type. Similar studies suggest that the sensorimotor cortex organizes grasp and object features in overlapping but distinct regions, modulated by task demands, object features, and the stage of the action (e.g., observation, preparation, grasping, and release) [9]. Furthermore, neural representations of the hand tend to reflect coordinated patterns of finger use in functional tasks rather than isolated joint movements [10, 11].

The present study aims to see whether artificial neural networks can develop intrinsic representations that mirror how the human brain links grasp types to object properties. Specifically, we ask whether such structure can emerge naturally in the model's latent space as a result of training. If successful, this would demonstrate a functional synergy between learned representations and biologically inspired affordance encoding. The contribution of such an understanding would be more efficient computing that is required for the real-time processing of the environment by embodied artificial cognitive systems (i.e., robots). Given that human-like hands have a high degree of freedom, and are capable of numerous different types of fine motor action, endowing robots with the mechanisms to plan the motor action in response to the environment in this abstract latent space instead of lower levels such as pixel, and single joint units, will significantly enhance such robots' capabilities for efficient and real-time operation.

2 Methodology

2.1 Dataset

In the current study, we utilized the HO-3D_v3 dataset[12], which contains 77,558 RGB frames annotated with precise 3D hand and object pose information during grasping interactions. This dataset includes 10 human subjects and 9 objects taken from the YCB dataset[13]. The YCB dataset is a standardized set of common household objects widely used in robotic grasping research due to its diversity, availability, and clearly defined physical properties and annotations. Each object is accompanied by high-resolution RGB-D (color and depth) frames and corresponding 3D annotations, such as hand pose, joint positions, object position, rotation, and dense point clouds. Objects include bleach cleaner, banana, meat can, scissors, cracker box, power drill, sugar box, mustard bottle, and mug. Annotations are structured into per-frame dictionaries that are available upon request for research purposes. The MANO human hand model [14] is adopted to represent hand poses in the dataset which facilitates further data-driven methods for pose estimation.

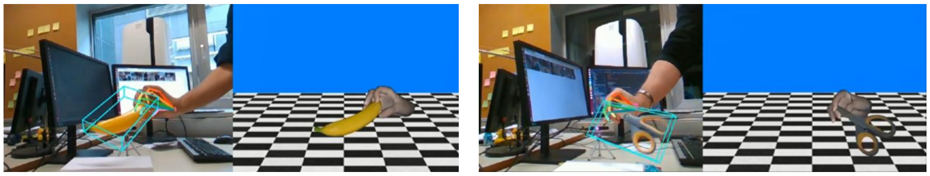
The MANO hand model includes 45 degrees of freedom (DoF) that consists of 15 overall finger joints with 3 DoF for each joint, and 6 DoF for the translation and rotation of the wrist which is the parent node of the kinematic tree of the hand. See Table1 for detailed hand-object information that we used from the dataset to train the CVAE model. Data preprocessing included feature extraction, normalization (via StandardScaler), and splitting into train-validation-test sets. The resulting dataset has 83,325 instances, each consisting of a concatenated 7983-dimensional feature vector representing hand-object pairs at each timestep. The HO-3D_v3 dataset was specifically selected because it provides precise, standardized annotations of diverse grasp-object interactions across multiple human subjects and object categories, minimizing potential confounds or biases that could lead to artefacts in our latent space analysis.

Data		**Dimensions**	
Hand	handPose	48	$48 + 3 + 63 = 114$
	handTrans	3	
	handJoints3D	$21 \times 3 = 63$	
Object	objectTrans	3	$3 + 3 + 7863 = 7869$
	objectRot	3	
	objectPointCloud	$2621 \times 3 = 7863$	
Hand Object Data			$114 + 7869 = 7983$

Table 1. Dimensional breakdown of hand and object features taken from HO-3D_v3 dataset

2.2 Conditional Variational Autoencoder (CVAE)

A Conditional Variational Autoencoder (CVAE) architecture was designed and trained to reconstruct 3D hand poses from object information for each grasping scenario. The model was inspired by previous works that have used a similar approach to generate grasping poses using CVAEs [15–17]. The network consists of an Object Encoder, a Hand Encoder, Prior and Posterior networks, and a



a) hand pose annotated based on banana

b) hand pose annotated based on scissors

Fig. 1. Examples from the HO-3D_v3 dataset showing hand-object interactions with (a) banana (011) and (b) scissors (037)[12]. The similar side grip in both cases may explain the close proximity of these objects in the PCA latent space despite their differing geometries.

Decoder to reconstruct the hand. The object encoder produces a 128D embedding from the object 6D information and its point cloud, while the hand encoder creates a 64D embedding from its 3D joint information, and global rotation and translation. The latent space is 64D, learned via reparameterization from the posterior conditioned on both hand and object encodings, while the prior is conditioned only on the object. A latent dimensionality of 64 was selected based on common practice in prior CVAE studies of grasp pose reconstruction, balancing representational capacity with computational efficiency. This dimensionality ensures that the latent space captures sufficient complexity to support meaningful PCA analysis, while avoiding excessive dimensionality that could lead to redundancy or instability in subsequent analyses.

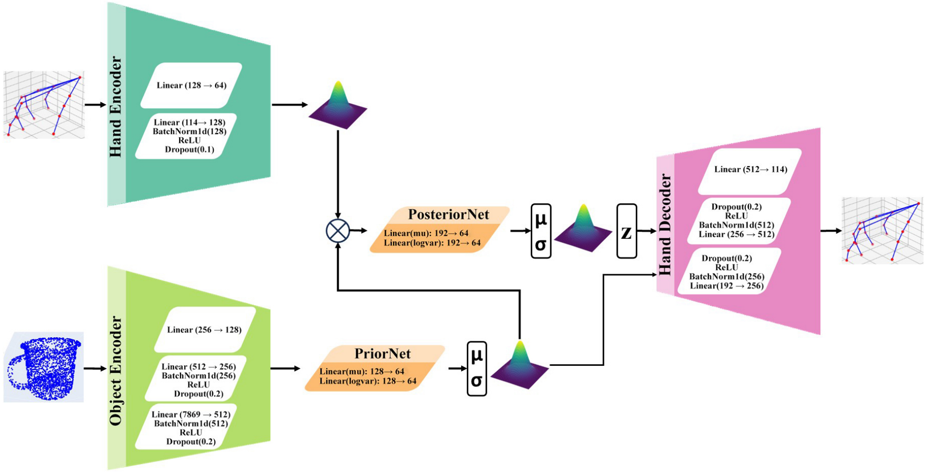


Fig. 2. Architecture of the Conditional Variational Autoencoder (CVAE) to reconstruct 3D hand poses conditioned on object features. The model consists of two encoders for the object and hand, a prior and a posterior network and a decoder to reconstruct the hand pose

The decoder reconstructs the hand from latent vectors concatenated with object embeddings. Dropout and batch normalization are applied throughout the model for regularization and training stability. Optimization uses the Adam optimizer with learning rate scheduling and early stopping based on validation loss. See Figure 2 for a more detailed model architecture.

The loss function used to train the CVAE is a weighted combination of the reconstruction loss and the Kullback–Leibler (KL) divergence regularization term. The reconstruction loss (Eq. 1) minimizes the difference between the predicted and ground-truth hand poses. The KL divergence (Eq. 2) regularizes the latent distribution to match a standard Gaussian distribution. The total loss (Eq. 3) is weighted by an annealing coefficient β, which gradually increases (Eq. 4).

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \tag{1}$$

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{j=1}^{d} \left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right) \tag{2}$$

$$\mathcal{L}_{\text{CVAE}} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathcal{L}_{\text{KL}} \tag{3}$$

$$\beta = \min\left(1, \frac{t}{T}\right) \tag{4}$$

Here, t is the current training epoch, and T is the total number of epochs over which the KL weight is gradually increased.

2.3 Latent Space Analysis

After training the CVAE model, we analyzed the learned latent representations using Principal Component Analysis (PCA). PCA reduces dimensionality and highlights the primary sources of variance within the latent space, allowing us to assess whether the learned representations meaningfully reflect grasp-object relationships. The latent vectors (64-dimensional) from the test set reconstructions were first extracted and subjected to PCA. We visualized the resulting lower-dimensional embeddings to examine the clustering patterns and determine whether they corresponded systematically to different objects or grasp types. Additionally, we computed cluster centroids for each object category and decoded these centroids back into 3D hand poses. This step allowed us to qualitatively verify if the latent space organization aligns with known grasp-object affordances observed in human sensorimotor studies.

3 Results

3.1 CVAE Training Performance

The CVAE model was trained on the HO-3D_v3 dataset using paired hand-object features while using object embeddings as conditional priors to guide the reconstruction of hand poses by the decoder. Training Loss, Validation Loss, and Reconstruction loss calculated as Mean Squared Error (MSE), decreased steadily during training, indicating that the model successfully learned to generate hand poses based on the presented object geometry. Early stopping was applied to prevent overfitting, and the decreasing loss trends confirm effective convergence. (See figure 4 for more details).

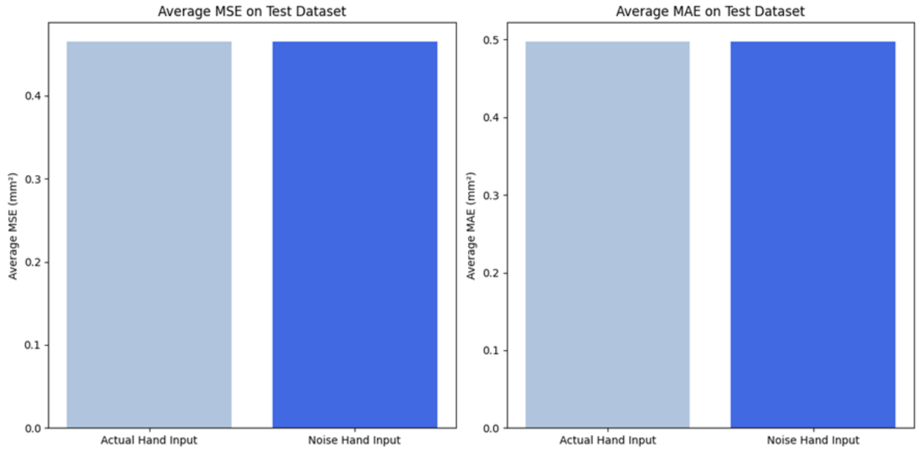


Fig. 3. Comparison of test-time reconstruction accuracy using posterior sampling (real hand input) versus prior sampling (Gaussian noise). Mean Squared Error (right) and Mean Absolute Error (left) yield comparable results for both conditions.

To validate the model's generalization capacity, hand pose reconstructions were tested on unseen data from the test set. Two strategies were employed: one using the real hand embedding (posterior sampling) and another using randomly sampled latent vectors from the learned prior distribution (prior sampling). Both approaches produced coherent hand pose outputs, with comparable Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics confirming low reconstruction error in both cases (See figure 3).

3.2 Principal Component Analysis (PCA)

To explore the internal organization of the learned latent space, Principal Component Analysis (PCA) was applied to the 64-dimensional latent vectors generated from the test set. The result as shown in Table 2 demonstrates that the first eight Principal Components captured 97.20% of the variance, with the first five PCs explaining 91.52% of all the variance, indicating that the latent space is effectively structured. This is in line with previous studies that suggest a few PCs are enough to represent grasping poses with acceptable accuracy [14, 18, 19].

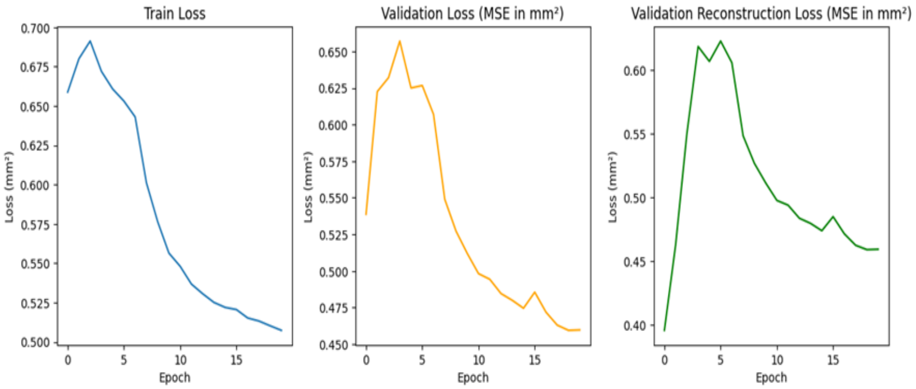
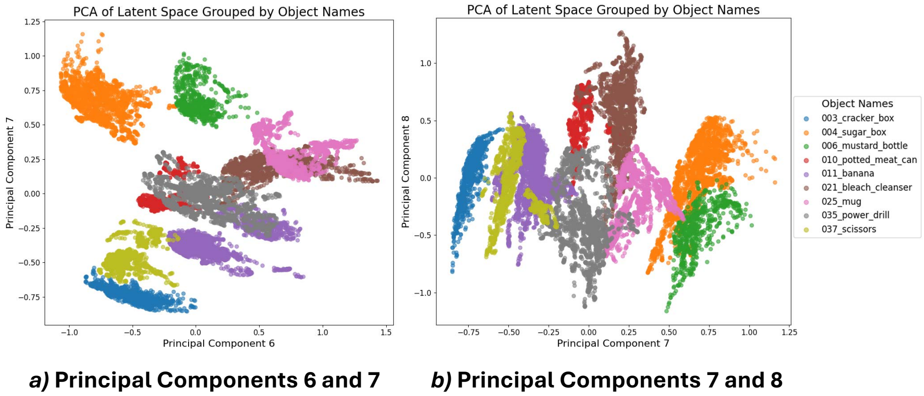


Fig. 4. Training curves showing CVAE optimization progress over epochs for Train Loss (left), Validation Loss (middle), and Reconstruction Loss (right)



***a)* Principal Components 6 and 7**

***b)* Principal Components 7 and 8**

Fig. 5. 2D PCA visualizations of the learned 64D latent space projected onto Principal Components (PCs) 6 & 7 (left) and 7 & 8 (right). Object-specific clusters emerge clearly in these subspaces, suggesting that the CVAE learns affordance-relevant structure that differentiates hand poses by object category.

Principle Component	Accumulative Explained Variance (%)
1	51.68
2	68.09
3	80.84
4	86.26
5	91.52
6	94.02
7	95.60
8	97.20
9	98.23
10	99.00

Table 2. Explained variance ratio of the first 10 principal components.

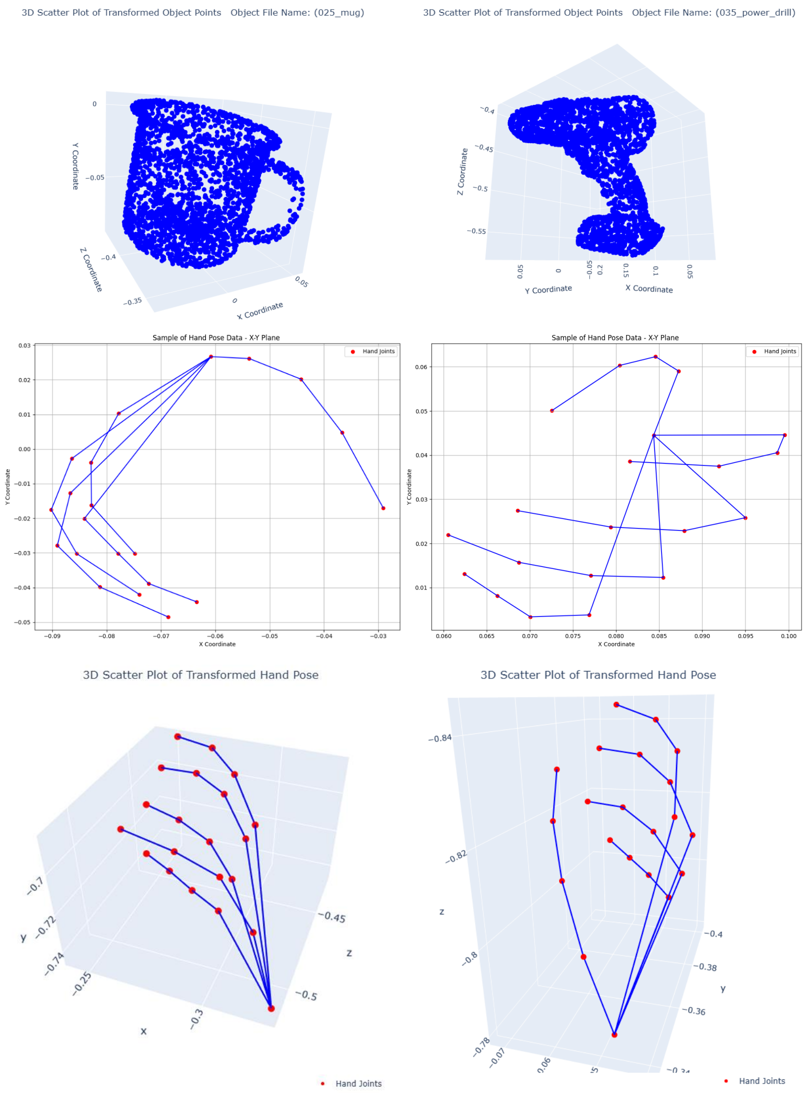
3.3 PCA Visualization and Clustering

Analysis of the PCA embeddings revealed that object identity–related clustering becomes prominent in Principal Components 5 through 8, as illustrated in Figure 5. Notably, while the first four components account for 86.26% of the total variance, they appear to capture global hand configuration features such as position and translation. In contrast, the emergence of object-specific clusters in later components suggests that the CVAE has implicitly organized affordance-relevant information within distinct subspaces of the latent space, which are more noticeable beyond the dominant variance dimensions.

In the Figure 5, the latent embeddings for ‘scissors’ and ‘banana’ are found in close proximity in the PCA space. This could be because of the similarity in the grasping postures supported—both objects commonly afford a parallel or side grasp with extended fingers. Although their geometric shapes are different, the resultant hand postures from the interaction have biomechanical similarities (Figure 1), which is why they overlap in the latent space. Two representative examples can be seen in Figure 6 where the centroids of clusters for objects mug and power drill are decoded back to the 3D pose using the decoder network. These findings suggest the emergence of an affordance-sensitive structure within the model, reminiscent of how the human brain encodes object-grasp relations across sensorimotor cortices.

4 Discussion

In this study, we trained a Conditional Variational Autoencoder (CVAE) to reconstruct human hand poses for grasping, conditioned on the object’s visual features. Once the model demonstrated robust training performance, we investigated the organization of its latent space using Principal Component Analysis (PCA). The analysis revealed that 5 to 8 principal components were sufficient to account for nearly all of the variance within the latent space, indicating that the model effectively learned a compact and structured internal representation.



***a)* Mug**

***b)* Power Drill**

Fig. 6. Reconstruction of decoded centroids from two latent space clusters corresponding to the objects “mug” (a) and “power drill” (b). These representative poses illustrate that distinct grasp types have emerged from abstract latent encodings, reflecting object-specific affordances.

Interestingly, we observed that some of the later components—specifically PCs 6 through 8—showed an explicit correspondence to object categories that afford specific types of grasps. This emergent clustering suggests a meaningful latent encoding that reflects affordance-based differentiation, echoing—albeit in a simplified form—the way the human brain encodes object-grasp associations in sensorimotor cortices. However, we acknowledge the possibility that the observed clustering may also be influenced by variation in hand poses rather than object features alone. Since our model reconstructs hand configurations conditioned on object encodings, the latent space inherently reflects this motor response. A valuable control would involve using a dataset in which a fixed grasp pose is applied across all objects. If clustering disappears under those conditions, it would suggest that pose diversity plays a central role in shaping the latent structure. Future work will explore this direction to better disentangle the contributions of object and hand features in affordance-based representation.

A particularly intriguing aspect of this finding is that such affordance-specific separability was not observed in the first few principal components. One hypothesis is that the early components are dominated by more general features of hand representation, such as global translation and wrist orientation, which capture broad variance but not task-specific structure. In contrast, the later components may encode more subtle, functionally relevant details concerning the relationship between the hand and the object, consistent with the functional encoding strategies observed in biological systems. A recent study [20] that introduced a framework to extract muscle synergies from a realistic model hand, found that the 8th synergy was most accountable for task execution, although it was not the most variable one. This suggests that key motor features may reside in mid-to-late latent dimensions. The agreement seen between our results and those of other researchers supports the hypothesis that both artificial and biological systems may rely on structured, lower-variance features to represent functionally important behaviors like grasp planning.

These findings offer promising implications for robotics. By leveraging such structured latent representations, robotic systems could achieve more efficient and generalizable control over high-degree-of-freedom manipulators. Rather than relying on low-level control at the pixel or joint level, abstract latent spaces may serve as higher-level control priors that facilitate real-time motor planning and execution.

Future work will extend this investigation to analyze the latent space in greater depth, with the aim of isolating more interpretable dimensions tied to intrinsic object properties such as shape, size, and orientation. Furthermore, we will explore whether these affordance-centric representations can be generalized to support motor planning for novel, previously unseen objects in unstructured environments—paving the way toward more adaptive and cognitively inspired robotic manipulation.

Acknowledgments. This work was funded by the Innovate UK (grant number 10089807 for the Horizon Europe project PRIMI Grant agreement n. 101120727). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC

BY) license to any Author Accepted Manuscript version of this paper arising from this submission.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. L. Shapiro, *Embodied cognition.* Routledge, 2019.
2. M. Wilson, "Six views of embodied cognition," *Psychonomic bulletin & review*, vol. 9, pp. 625–636, 2002.
3. R. A. Zwaan, "Two challenges to "embodied cognition" research and how to overcome them," *Journal of Cognition*, vol. 4, no. 1, p. 14, 2021.
4. J. J. Gibson, "The theory of affordances:(1979)," in *The people, place, and space reader.* Routledge, 2014, pp. 56–60.
5. E. Cesanek, Z. Zhang, J. N. Ingram, D. M. Wolpert, and J. R. Flanagan, "Motor memories of object dynamics are categorically organized," *Elife*, vol. 10, p. e71627, 2021.
6. S. J. Anderson, N. Yamagishi, and V. Karavia, "Attentional processes link perception and action," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 269, no. 1497, pp. 1225–1232, 2002.
7. J. Grèzes, M. Tucker, J. Armony, R. Ellis, and R. E. Passingham, "Objects automatically potentiate action: an fmri study of implicit processing," *European Journal of Neuroscience*, vol. 17, no. 12, pp. 2735–2740, 2003.
8. A. I. Sburlea, M. Wilding, and G. R. Müller-Putz, "Disentangling human grasping type from the object's intrinsic properties using low-frequency eeg signals," *Neuroimage: Reports*, vol. 1, no. 2, p. 100012, 2021.
9. S. Fabbri, K. M. Stubbs, R. Cusack, and J. C. Culham, "Disentangling representations of object and grasp properties in the human brain," *Journal of Neuroscience*, vol. 36, no. 29, pp. 7648–7662, 2016.
10. N. Ejaz, M. Hamada, and J. Diedrichsen, "Hand use predicts the structure of representations in sensorimotor cortex," *Nature neuroscience*, vol. 18, no. 7, pp. 1034–1040, 2015.
11. U. Castiello, "The neuroscience of grasping," *Nature Reviews Neuroscience*, vol. 6, no. 9, pp. 726–736, 2005.
12. S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3196–3206.
13. T. Grenzdörffer, M. Günther, and J. Hertzberg, "Ycb-m: A multi-camera rgb-d dataset for object recognition and 6dof pose estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 2020, pp. 3650–3656.
14. J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.
15. J. Hang, X. Lin, T. Zhu, X. Li, R. Wu, X. Ma, and Y. Sun, "Dexfuncgrasp: A robotic dexterous functional grasp dataset constructed from a cost-effective real-simulation annotation system," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 9, 2024, pp. 10 306–10 313.

16. Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu *et al.*, "Realdex: Towards human-like grasping for robotic dexterous hand," *arXiv preprint arXiv:2402.13853*, 2024.
17. H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 107–11 116.
18. Q. Lei, G. Chen, and M. Wisse, "Fast grasping of unknown objects using principal component analysis," *Aip Advances*, vol. 7, no. 9, 2017.
19. C. Rolinat, M. Grossard, S. Aloui, and C. Godin, "Learning to model the grasp space of an underactuated robot gripper using variational autoencoder," *IFAC-PapersOnLine*, vol. 54, no. 7, pp. 523–528, 2021.
20. C. Berg, V. Caggiano, and V. Kumar, "Sar: Generalization of physiological agility and dexterity via synergistic action representation," *Autonomous Robots*, vol. 48, no. 8, p. 28, 2024.