Large Behavior Models: a promising approach for safe and reliable robotics?

Branislav Zigo, Igor Farkaš

Department of Applied Informatics, Comenius University Bratislava
{branislav.zigo, igor.farkas}@fmph.uniba.sk

Abstract

Vision-Language-Action models (VLAs) and Large Behavior Models (LBMs) are becoming a central research direction for cognitive and humanoid robotics. They promise to connect semantic perception, natural-language task specification, multimodal state estimation, and continuous motor control in one trainable policy stack. This paper reviews the motivation, historical roots, architectural principles, and taxonomy of recent VLA and LBM approaches. Particular attention is given to multimodal and multisensory fusion, cross-embedding of vision, language, robot state and action, autoregressive action token generation, diffusion policy, and flow matching. The paper also discusses robot datasets, representative systems and prediction-based alternatives. The main conclusion is that cognitive robotics is moving toward hybrid architectures that combine pretrained semantic backbones, multisensory latent state estimation, generative continuous action heads and predictive world models. In the end, we also discuss the limitations of this approach to building intelligent robotic systems.

1 Introduction

Humanoid robots are no longer only mechanical platforms for scripted motion. They are becoming embodied AI systems that must interpret human instructions, perceive complex scenes, interact safely with objects and people, and coordinate many degrees of freedom in real time. This requirement creates a difficult cognitive-robotics problem: the robot must connect symbolic or linguistic task descriptions with perceptual grounding, physical state estimation, and closed-loop control. Vision-Language-Action models (VLAs) address this problem by extending Vision-Language Models (VLMs) toward direct robot action generation (Kawaharazuka et al., 2025). Large Behavior Models (LBMs) address a broader problem: how to scale robot policies across many behaviors, tasks, environments and embodiments (TRI LBM Team et al., 2026).

The distinction between VLA and LBM is useful but not absolute. A VLA is usually language-grounded: it receives an image or video observation, a language instruction, and possibly proprioceptive state, then produces actions (see an example in Fig.1). An LBM is a large-scale behavior policy trained on broad robot data; it may be language-conditioned, but it can also be defined primarily by action competence and data scale. The strongest current robot foundation models increasingly combine semantic grounding from pretrained vision-language backbones, broad behavioral competence from large trajectory datasets, and continuous action generation for physical control (Octo Model Team et al., 2024; Black et al., 2024; NVIDIA et al., 2025).

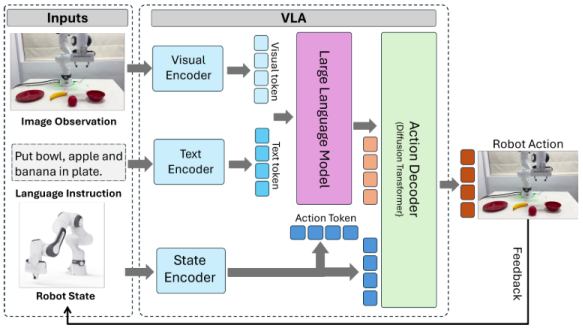

Fig. 1: Architecture of a typical Vision-Language-Action model for robotic manipulation (Din et al., 2025). The three embeddings are fused in an LLM that generates multimodal semantic representation of the intended task, i.e., along with robot state features, decoded to a trajectory that accomplishes the task.

For cognitive and humanoid robotics, this shift is important because classical modular pipelines often separate perception, mapping, planning, grasp synthesis and motor control. Such modularity remains valuable for verification and safety, but it can be brittle under open-ended instructions, clutter, occlusion, deformable objects, contact-rich manipulation and embodiment mismatch. VLA and LBM research does not remove the need for structure; rather, it asks how much of this structure can be learned, shared across tasks and adapted to new robots.

There also exist recent robotics-related proto-LBMs that stand somewhere between VLAs and LBMs. RoboVLM (Li et al., 2026) and HY-Embodied (Yu et al., 2026) differ in details but converge on the same idea: a unified, scalable VLA foundation model that generates behavior directly, an emerging form of LBMs.

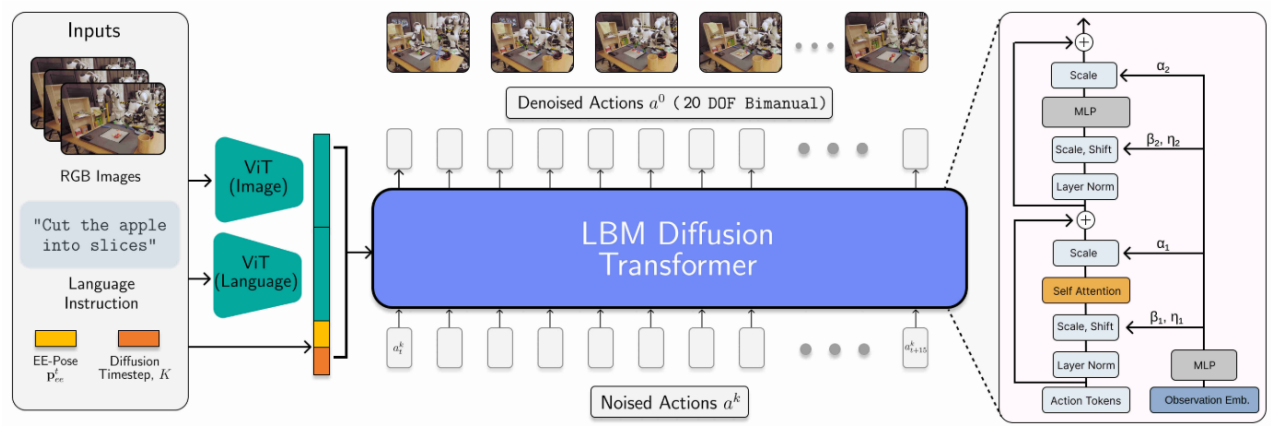

Fig. 2: Architecture of a Large Behavior Model (TRI LBM Team et al., 2026) The model is conditioned on language, vision and proprioception and outputs 20-dimensional actions over 16 timesteps.

2 Historical roots and motivation

The first root of VLAs and LBMs is end-to-end visuomotor learning. Early deep visuomotor policies showed that perception and control could be trained jointly from raw sensory observations, challenging the idea that a robot always needs a manually specified symbolic state before action selection (Noda et al., 2014). RT-1 later demonstrated that transformer architectures can absorb large real-world robot datasets and generalize across many tasks when camera observations, instructions and motor commands are represented in a compact sequence format (Brohan et al., 2022).

The second root is multimodal representation learning that studies how models represent, align, translate, fuse and co-learn from multiple modalities (Baltrušaitis et al., 2019). Early deep multimodal learning showed that coordinated representations can transfer information across sensory streams (Ngiam et al., 2011). In robotics, multimodal integration was already studied on humanoid platforms by learning fused sensorymotor time-series representations for behavior generation (Noda et al., 2014). This tradition becomes essential in manipulation, where vision may see the object globally, proprioception tells the robot where its body is, and tactile or force feedback reveals contact that cameras cannot observe (Li et al., 2023; Sferrazza et al., 2024).

The third root is language-conditioned imitation learning and vision-language pretraining. CLIP showed that large-scale image-text pretraining can produce transferable visual concepts addressable through language (Radford et al., 2021). CLIPort then combined CLIP-like semantic understanding with spatial manipulation precision, separating the question of what to manipulate from where to act (Shridhar et al., 2022). PaLM-E inserted continuous embodied observations into the token stream of a large language model (LLM), showing that language, image and robot state can be processed jointly for embodied reasoning (Driess et al., 2023). VIMA generalized the prompting view by interleaving visual and textual prompt tokens and decoding robot actions autoregressively (Jiang et al., 2022).

The final root is sequence modelling for decision making. Decision Transformer and Trajectory Transformer reframed control as sequence prediction over states, actions and returns (Chen et al., 2021; Janner et al., 2021). This was conceptually important because it made robot control compatible with the scaling recipes of LLMs: tokenize inputs, train on large sequential datasets, and predict the next element in context. RT-2 made this connection explicit by expressing robot actions as text-like tokens and co-training on robot trajectories and web-scale vision-language tasks (Zitkovich et al., 2023).

3 Key architectural principles

3.1 Multimodal and multisensory fusion

Multimodal fusion means that a policy does not rely on one input channel only. In VLAs and LBMs, the usual modalities are image or video, language instruction, robot proprioception, action history and sometimes tactile, force, audio or depth observations (Baltrušaitis et al., 2019). Multisensory fusion is especially important for humanoid manipulation because many task-relevant variables are hidden: the object may be occluded by the hand, contact pressure may change before the object visibly moves, and slippage may be felt before it is seen. Experiments on visual-auditory-tactile manipulation show that different senses dominate different parts of the task: vision supplies global scene context, audio can mark contact events, and touch supplies local geometry and contact information (Li et al., 2023). Vision-touch masked multimodal learning similarly shows that tactile information improves sample efficiency and robustness under occlusion (Sferrazza

et al., 2024).

For cognitive robotics, the key point is that the robot internal state should not be a simple image caption. It should be a control-relevant latent state: a compact representation that binds object identity, spatial relations, affordances, body configuration, contact state and task intention. Here the cross-embedding becomes central.

3.2 Cross-embedding and embodied tokens

Cross-embedding refers to representing multiple modalities in a common or coordinated latent space so that information can be compared, attended to, fused or translated across modalities (Ngiam et al., 2011; Baltrušaitis et al., 2019). In current robot foundation models, this often means mapping visual features, language tokens and robot state vectors into a transformer-compatible sequence. PaLM-E is a clear example: visual observations and continuous state estimates are projected into the same processing stream as language tokens (Driess et al., 2023). VIMA uses multimodal prompts in which text and visual tokens specify the manipulation task (Jiang et al., 2022). OpenVLA combines a Llama-based decoder with fused DINOv2 and SigLIP visual features, illustrating how pretrained visual and vision-language encoders can support robot action prediction (Kim et al., 2025; Oquab et al., 2023; Zhai et al., 2023).

This design has a cognitive interpretation. The robot does not merely classify pixels or parse commands. It constructs an embodied context in which words such as cup, fragile, left, gently or give become linked to visual objects, body posture, force constraints and possible action sequences. Cross-embedding is therefore a computational mechanism for grounding meaning in action.

3.3 Autoregressive token generation

Autoregressive token generation means that the model predicts the next token conditioned on previous tokens. In language models the tokens are words or subwords; in robot policies they may be discretized actions, action chunks or symbolic representations of continuous control variables (Chen et al., 2021). RT-2 is the canonical VLA example: it formats robot actions as text-like tokens so that the same model can learn both language responses and action outputs (Zitkovich et al., 2023). OpenVLA keeps this transformer-based VLA idea but releases an open model and training stack, making the approach more reproducible (Kim et al., 2025).

The advantage of autoregressive VLAs is conceptual and engineering simplicity. They fit naturally with pretrained VLMs and allow web-scale semantic knowledge to influence robot control. Their limitation is action precision. Fine motor control often requires smooth, high-frequency, continuous trajectories. Discretizing a dexterous hand, wrist and arm into tokens can introduce quantization error, longer sequences and latency. This limitation motivates diffusion and flow-based action heads.

4 VLA and LBM model families

A practical taxonomy can be organized by the way task context and action are represented.

(1) Semantic-spatial manipulation models use vision-language pretraining to ground object concepts while preserving precise spatial control. CLIPort is the representative example because it combines a semantic pathway for what to manipulate with a spatial pathway for where to manipulate (Shridhar et al., 2022). Such models are not always full VLAs in the recent foundation-model sense, but they are historically important because they demonstrated that internet-scale semantic representations can improve robot manipulation.

(2) Promptable and embodied language models process text, visual observations and sometimes robot state in a common sequence. PaLM-E is primarily an embodied multimodal language model for reasoning and planning (Driess et al., 2023), whereas VIMA is a transformer agent that solves tabletop tasks from multimodal prompts and outputs actions autoregressively (Jiang et al., 2022). These models make the instruction interface explicit and show how language can structure robot behavior beyond fixed task labels.

(3) Autoregressive VLA controllers directly treat action as part of the token sequence. RT-1 introduced a scalable robotics transformer trained on diverse real-world robot data (Brohan et al., 2022). RT-2 then transferred web-scale vision-language knowledge into robot control by co-training on internet vision-language tasks and robot trajectories (Zitkovich et al., 2023). OpenVLA extended this line as an open 7B-parameter VLA trained on 970k robot demonstrations from Open X-Embodiment (Kim et al., 2025; O'Neill et al., 2023).

(4) Generalist robot policies and LBMs emphasize behavioral scale across tasks and embodiments. Octo is a transformer-based diffusion policy trained on 800k trajectories from Open X-Embodiment and designed for adaptation to new sensors and action spaces (Octo Model Team et al., 2024; O'Neill et al., 2023). TRI's LBMs extend the diffusion-policy paradigm across approximately 1,700 hours of robot data and evaluate multitask dexterous manipulation with controlled real-world and simulation trials (TRI LBM Team et al., 2026). In this family, language is useful, but the defining feature is broad behavioral competence from large robot data.

(5) Continuous generative policies use diffusion or flow matching to produce smooth action trajectories. Diffusion Policy models actions as a denoising process conditioned on observations (Chi et al., 2025). RDT-1B scales a diffusion transformer for bimanual manipula-

tion (Liu et al., 2024). GR00T N1 uses a dual-system VLA architecture for humanoids in which a vision-language module interprets the instruction and a diffusion transformer generates motor actions (NVIDIA et al., 2025). The π_0 model uses flow matching rather than classical action tokenization to generate continuous actions from a pretrained vision-language backbone (Black et al., 2024).

(6) Predictive world-model alternatives such as JEPA-based systems learn how the world evolves in latent space and then plan actions, instead of directly mapping observation and instruction to action (Assran et al., 2025; Terver et al., 2025). They are not merely another action head; they represent a different hypothesis about cognitive robotics: robust behavior may require prediction and planning in an abstract world model.

5 Datasets and scaling

The scaling of VLAs and LBMs depends on datasets as much as architectures. Open X-Embodiment pooled robot trajectories across many institutions and robot embodiments, enabling the RT-X line of cross-robot policies and providing a foundation for OpenVLA and Octo (O'Neill et al., 2023; Kim et al., 2025; Octo Model Team et al., 2024). DROID extended the dataset landscape with in-the-wild robot manipulation demonstrations collected across many real scenes, tasks and data collectors (Khazatsky et al., 2024). VIMA-Bench, introduced with VIMA, provides procedurally generated tabletop tasks with multimodal prompts, supporting systematic evaluation of generalization (Jiang et al., 2022). Another interesting dataset is Humanoid Everyday that aims to advance research in general-purpose humanoid manipulation and lay the groundwork for more capable and embodied robotic agents in real-world scenarios (Zhao et al., 2025).

However, robot data remain scarce compared with internet text and image data. This is the central bottleneck for humanoid robotics. Humanoids have high-dimensional bodies, many coupled joints, unstable contacts and safety constraints. Consequently, near-future VLA and LBM training will likely combine robot teleoperation, simulation, motion capture, human video, synthetic data and self-supervised world-model pretraining (O'Neill et al., 2023; Khazatsky et al., 2024; NVIDIA et al., 2025; Assran et al., 2025).

6 Diffusion policy and flow matching

6.1 Diffusion policy

Diffusion policy applies the idea of denoising diffusion models to robot action generation (Chi et al., 2025). During training, the model sees expert action trajectories, adds different levels of noise to them, and learns how to remove that noise while conditioning on observations such as images, robot state and task context. During inference, it begins with a noisy action sequence and iteratively denoises it into a plausible action trajectory. This is useful because manipulation is often multimodal: there may be several correct ways to grasp a cup, fold a cloth or move around an obstacle. A standard regression policy may average these alternatives into a bad action, while a diffusion policy can represent several possible action modes (Chi et al., 2025).

In robotics, diffusion policies are often used with receding-horizon control (Chi et al., 2025). The policy generates an action chunk, the robot executes the first part, receives updated observations, and then replans. This provides a compromise between long-horizon temporal consistency and closed-loop correction. Octo, RDT-1B, GR00T N1 and TRI LBMs can all be read as different ways of scaling this continuous generative-action idea to larger datasets, broader tasks or embodiments (Octo Model Team et al., 2024; Liu et al., 2024; NVIDIA et al., 2025; TRI LBM Team et al., 2026).

The price of diffusion is sampling cost. Since the action is refined through multiple denoising steps, inference may be slower than one-shot prediction. This matters for dexterous hands, whole-body humanoids and dynamic interaction. Research therefore explores fewer denoising steps, action chunking, specialized diffusion transformers and alternatives such as flow matching (Lipman et al., 2023; Black et al., 2024).

6.2 Flow matching and the velocity-field analogy

Flow matching is a generative modelling objective that learns a vector field transforming simple noise into samples from the data distribution (Lipman et al., 2023). The intuitive analogy is a river current. Imagine many small floating leaves initially scattered randomly on the water. A velocity field tells each leaf in which direction and how fast to move at each point. If the current is learned correctly, the random cloud of leaves gradually flows into the shape of the desired object. In a robot policy, the random cloud is initial action noise, the final shape is a meaningful action trajectory, and the learned current is the policy's velocity field (Black et al., 2024).

Compared with classical diffusion, flow matching can be understood as learning a more direct transport path from noise to data. It does not only ask the model to remove noise step by step; it asks the model to learn the instantaneous direction of travel along a probability path. The π_0 model uses this principle to generate continuous robot actions with a separate action expert on top of a pretrained vision-language backbone (Black et al., 2024). This is attractive for fine motor control because actions remain continuous, action chunks can be smooth, and the semantic backbone can still provide language and visual grounding.

For an audience new to the velocity-field concept,

a second analogy is navigation in a city. A map tells where the streets are, but a velocity field tells how traffic should move at every location and time. Flow matching trains the policy to produce this motion rule. Once trained, the model can start from an unstructured initial action and follow the learned flow toward a coherent motor command. For humanoid robotics, this is promising because the same high-level instruction may require subtle continuous coordination of shoulder, elbow, wrist and fingers.

7 JEPA and predictive world-models

Joint-Embedding Predictive Architectures (JEPAs) offer a different route toward embodied intelligence. A JEPA predicts future or missing representations in latent space rather than reconstructing raw pixels (LeCun, 2022; Assran et al., 2023). I-JEPA demonstrated this principle for images (Assran et al., 2023), while V-JEPA 2 extends it to video and physical planning by learning from large-scale video and then adding action-conditioned latent dynamics for robotic tasks (Assran et al., 2025). JEPA-based world models aim to support planning by predicting how the world will change under possible actions, and recent work evaluates which architectural and training choices are important for physical planning in latent space (Terver et al., 2025).

The difference from direct VLA policy learning is fundamental. A VLA often maps observation and instruction directly to action. A JEPA-style world model first learns a predictive representation of the world, then uses planning to choose actions that move the latent state toward a goal (LeCun, 2022; Assran et al., 2025). This may be advantageous for long-horizon cognition, causal anticipation and learning from passive video, but it can add planning cost and currently has less mature language-to-action grounding than VLA controllers. A plausible future is hybrid: a VLA proposes grounded actions, while a predictive world model evaluates consequences and supports lookahead.

8 Implications for cognitive and humanoid robotics

The VLA-to-LBM transition changes how we can think about robot cognition. Cognition is no longer located only in symbolic planning or explicit world models. It is distributed across pretrained perceptual backbones, cross-modal embeddings, temporal sequence models, action generators and feedback loops. In humanoid robots, this matters because behavior depends on the whole body: gaze, torso, arms, hands, tactile contact and balance must cooperate. GR00T N1 and related humanoid foundation models explicitly target this whole-body setting (NVIDIA et al., 2025), while bimanual diffusion models such as RDT-1B target the coordination problem of two arms (Liu et al., 2024).

At the same time, several challenges remain open.
(1) Generalization is still limited by data coverage: a model can only generalize robustly if its training distribution contains enough variation in objects, scenes, embodiments and failures (O'Neill et al., 2023; Khazatsky et al., 2024; TRI LBM Team et al., 2026).
(2) Evaluation is difficult because robot success is stochastic, hardware-dependent and sensitive to small environmental changes (TRI LBM Team et al., 2026).
(3) Safety and interpretability are unresolved. In human environments, a humanoid should not only act successfully but also make its intentions understandable, detect uncertainty, and fail safely.
(4) Multimodal fusion must handle missing or unreliable sensors; a camera may be occluded, a tactile sensor may drift, and a language instruction may be ambiguous (Li et al., 2023; Sferrazza et al., 2024).

These limitations suggest that future cognitive humanoid systems will not be pure end-to-end black boxes. They will probably combine learned policies with constraints, safety monitors, symbolic task structure, uncertainty estimation, human feedback and predictive world models. The contribution of VLA and LBM research is that it supplies a scalable learned substrate on which such structured cognition can be built.

9 Core problems of MLLMs

In general, despite their interesting performance, MLLMs are subject to several shortcomings (which apply to VLAs and LBMs as well). One is the well-known problem of **"hallucinations"** (Ji et al., 2023), which was first identified in LLMs, primarily linked to the symbol grounding problem (SGP) Huang et al. (2023). Since its discovery, there has been a growing number of literature referring to this undesirable phenomenon manifesting itself in various forms (Alansari and Luqman, 2026). Kalai and Vempala (2024) show that, under calibration assumptions and for certain classes of arbitrary facts, hallucination must occur, i.e., it has a statistical lower bound rather than being only an engineering defect. Regarding models' internal representations, the existence of the so-called H-neurons responsible for hallucinations has been reported (Gao et al., 2025). From the psychological perspective, the hallucinating LLMs have been shown to cause delusional spiraling even in very rational users (Chandra et al., 2026). Obviously, the existence of ubiquitous hallucinations poses substantial obstacles to practical deployment of LLMs and raises concerns regarding their reliability in real-world applications.

However, the expectations that SGP can be solved by grounding symbols in robot experience using MLLMs have not been met, since the hallucinations

have been shown not to be language-specific (Sun et al., 2024). According to most recent literature, hallucinations have also been found in MLLMs Huang et al. (2025). Vision-Language Models (VLM) often exhibit cross-modal inconsistency, i.e. a discrepancy between an image content and text response. For instance, in case of object-related hallucinations (a frequent case), three types have been identified: object category, object attribute, and object relation. Hallucinating MLLMs are a problem which implies that merely adding modalities to language, in order to address the SGP and achieve understanding, does not automatically provide a solution. The core of the problem must lie elsewhere. For instance, Farkaš et al. (2025) argue that the problem may be due to a fundamentally different (non-developmental) way how world knowledge is acquired in MLLMs compared to humans, the main argument being that in MLLMs the language is central and that they lack non-linguistic world model (unlike people).

Apart from hallucinations, another problem of MLLMs has been identified, labeled **"mirage effect"** as evidence of an illusion of visual understanding in VLMs (Asadi et al., 2026). The authors define mirage as an AI model describing visual features in the absence of any image input. The output of a model affected by the mirage effect, seen in isolation, is indistinguishable from that of normal visual reasoning. The model does this without expressing any uncertainty, lack of confidence, or acknowledging an assumption or hypothetical scenario. The authors explain that contrary to hallucinations, the mirage effect does not necessarily involve inconsistencies or false responses. The response in mirage-mode can be correct in every sense, accompanied by a meticulous reasoning trace, and completely coherent. The main characteristic of the mirage effect, however, is the construction of a false epistemic frame that is not grounded in the provided input. Mirage effect is hence also a problem that hinders the practical deployment of such VLM models.

10 Challenges for LBMs in robotics

Despite growing excitement about the potential of leveraging AI in robotics, it must be acknowledged perceving and acting in the physical world pose greater and different challenges for AI than analysing data in isolation (Billard et al., 2025). More concretely, we cannot expect approaches that excel in purely data and software-based gaming environments, or image or text generation, to be readily applied to real-time sensing, planning, control and navigation for physical machines operating in complex, unpredictable environments, especially those involving human interaction. As a roadmap for AI in robotics, several short-term challenges have been identified (Billard et al., 2025): (1) Improved datasets (for manipulation and navigation tasks), for different tasks and different embodiments and interaction with humans, (2) Techniques to overcome the sim-to-real gap, (3) Combination of model-based and model-free learning, and (4) Generative models for robotics including the use of existing LLMs for semantic inference, development of new large-scale visual models and VLA models for different tasks.

The paper also mentions the long-term challenges, namely (1) life-long learning which would allow robots to acquire new knowledge and skills during their entire operational life, and (2) transfer learning which would enable transferring learned skills to new tasks, different robots and environments, perceptions between robots, or from simulation to physical robots. Given the expanding literature on VLAs, transfer learning is already being addressed intensively, supported by a growing number of large multimodal datasets.

11 Near-future outlook

In the near future, four developments seem likely.

(1) Continuous action generation will become more important for dexterous and humanoid robotics. Autoregressive token policies will remain attractive for semantic grounding, but diffusion and flow matching better match the smooth control needs of hands and whole-body motion (Chi et al., 2025; Black et al., 2024).

(2) Multisensory fusion will move from an optional add-on to a core requirement, especially for manipulation under occlusion, deformable objects and contact-rich tasks (Li et al., 2023; Sferrazza et al., 2024).

(3) Data mixtures will become increasingly heterogeneous. Robot-only data are too expensive, so training will combine robot trajectories, human demonstrations, egocentric video, simulation, synthetic data and self-supervised video pretraining (O'Neill et al., 2023; Khazatsky et al., 2024; Assran et al., 2025).

(4) The boundary between policy models and world models will blur. VLAs will gain predictive modules, and JEPA-style models will gain stronger language grounding and action interfaces (Assran et al., 2025; Terver et al., 2025).

For cognitive robotics, the most promising direction is therefore not a single architecture but a layered recipe: pretrained semantic perception, multisensory state embedding, generative continuous action, predictive latent planning, and safety-aware human interaction. This recipe can support robots that are not only more capable, but also more adaptive and understandable in everyday human environments.

12 Conclusion

VLAs and LBMs represent a shift from task-specific robotic programs toward scalable embodied policies. VLAs emphasize the grounding of language and vi-

sion in action, while LBMs emphasize broad behavioral competence across tasks and data. Their foundations lie in multimodal learning, language-conditioned imitation learning, sequence modelling and robot dataset scaling. Current state-of-the-art systems differ in how they represent actions: autoregressive token models connect naturally to language-model training; diffusion policies represent multimodal continuous trajectories; flow-matching policies learn a velocity field from noise to action; and JEPA-based world models provide a predictive alternative for planning in latent space.

For humanoid cognitive robotics, the central opportunity is integration. A useful humanoid robot must understand instructions, perceive objects, infer hidden physical state, coordinate the whole body, and adapt under uncertainty. No single model family solves all of these requirements. The likely future is hybrid: VLA-style semantic grounding, LBM-style behavioral scale, diffusion or flow-based motor generation, and JEPA-style prediction. Such systems could move humanoid robots from impressive demonstrations toward robust, explainable and socially useful behavior.

Acknowledgment: Supported by EU NextGenerationEU through Recovery and Resilience Plan for Slovakia, project HUROSI, number 09I01-03-V04-00048, and in part by Horizon Europe MSCA project TRAIL, number 101072488.

References

Alansari, A. and Luqman, H. (2026). Large language models hallucination: A comprehensive survey. arXiv:2510.06265.

Asadi, M., O'Sullivan, J. W., Cao, F., Nedaee, T., Rajabalifardi, K., Li, F.-F., Adeli, E., and Ashley, E. (2026). MIRAGE: The illusion of visual understanding. arXiv:2603.21687.

Assran, M., , et al. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv:2506.09985.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Conference on Computer Vision and Pattern Recognition*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Billard, A. et al. (2025). A roadmap for AI in robotics. *Nature Machine Intelligence*, 7:818–824.

Black, K. et al. (2024). π_0: A vision-language-action flow model for general robot control. arXiv:2410.24164.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. (2022). RT-1: Robotics transformer for real-world control at scale. arXiv:2212.06817.

Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., and Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal bayesians. arXiv:2602.19141.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, pages 15084–15097.

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. (2025). Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*.

Din, M. U., Akram, W., Rosell, L. S. S. J., and Hussain, I. (2025). Vision language action models in robotic manipulation: A systematic review. arXiv:2507.10672.

Driess, D. et al. (2023). PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488.

Farkaš, I., Vavrečka, M., and Wermter, S. (2025). Will multimodal large language models ever achieve deep understanding of the world? *Frontiers in Systems Neuroscience*, 19. doi:10.3389/fnsys.2025.1683133.

Gao, C., Chen, H., Xiao, C., Chen, Z., Liu, Z., and Sun, M. (2025). H-neurons: On the existence, impact, and origin of hallucination-associated neurons in LLMs. arXiv:2512.01797.

Huang, L. et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2). doi:10.1145/3703155.

Huang, S. et al. (2023). Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*.

Janner, M., Li, Q., and Levine, S. (2021). Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, pages 1273–1286.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung,

P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12). doi:10.1145/3571730.

Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., and Fan, L. (2022). VIMA: General robot manipulation with multimodal prompts. arXiv:2210.03094.

Kalai, A. T. and Vempala, S. S. (2024). Calibrated language models must hallucinate. In *Annual ACM Symposium on Theory of Computing*, page 160–171. doi:10.1145/3618260.3649777.

Kawaharazuka, K., Oh, J., Yamada, J., and Posner, I. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. doi:10.1109/ACCESS.2025.3609980.

Khazatsky, A. et al. (2024). DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*.

Kim, M. J. et al. (2025). OpenVLA: An open-source vision-language-action model. In *8th Conference on Robot Learning*, pages 2679–2713.

LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview. https://openreview.net/forum?id=BZ5a1r-kVsf.

Li, H. et al. (2023). See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Proceedings of Machine Learning Research*, pages 1368–1378.

Li, X. et al. (2026). What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, 8:158–172. doi:10.1038/s42256-025-01168-7.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *International Conference on Learning Representations*.

Liu, S. et al. (2024). RDT-1B: A diffusion foundation model for bimanual manipulation. arXiv:2410.07864.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696.

Noda, K., Arie, H., Suga, Y., and Ogata, T. (2014). Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems*, 62(6):721–736.

NVIDIA, Bjorck, J., and Zhu, Y. (2025). GR00T N1: An open foundation model for generalist humanoid robots. arXiv:2503.14734.

Octo Model Team et al. (2024). Octo: An open-source generalist robot policy. arXiv:2405.12213.

O'Neill, A. et al. (2023). Open X-embodiment: Robotic learning datasets and RT-X models. arXiv:2310.08864.

Oquab, M. et al. (2023). DINOv2: Learning robust visual features without supervision. arXiv:2304.07193.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of Machine Learning Research*, volume 139, pages 8748–8763.

Sferrazza, C., Seo, Y., Liu, H., Lee, Y., and Abbeel, P. (2024). The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning. In *International Conference on Intelligent Robots and Systems*, pages 9698–9705.

Shridhar, M., Manuelli, L., and Fox, D. (2022). Cliport: What and where pathways for robotic manipulation. In *Proceedings of Machine Learning Research*, volume 164, pages 894–906.

Sun, Y., Sheng, D., Zhou, Z., and Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*. doi:10.1057/s41599-024-03811-x.

Terver, B., Yang, T.-Y., Ponce, J., Bardes, A., and LeCun, Y. (2025). What drives success in physical planning with joint-embedding predictive world models? arXiv:2512.24497.

TRI LBM Team et al. (2026). A careful examination of large behavior models for multitask dexterous manipulation. *Science Robotics*. doi:10.1126/scirobotics.aea6201.

Yu, X. et al. (2026). HY-embodied-0.5: Embodied foundation models for real-world agents. arXiv:2604.07430.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pages 11941–11952.

Zhao, Z. et al. (2025). Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. arXiv:2510.08807.

Zitkovich, B. et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of Machine Learning Research*, volume 229, pages 2165–2183.