Adaptivní robotické systémy: Učení a přizpůsobování napříč úlohami a uživateli





Karla Štěpánová

Robotic Perception Group

Czech Technical University in Prague
Czech Institute of Informatics, Cybernetics and Robotics (CIIRC)
160 00 Prague 6, Jugoslávských partyzánů 1580/3, Czech Republic

karla.stepanova@cvut.cz, http://karlastepanova.cz

http://rop.ciirc.cvut.cz

Where we fit into CIIRC organization structure?

CIIRC consists of these scientific departments



1. Artificial Intelligence (Mikoláš Janota)
2. Intelligent systems (Vladimír Mařík)
3. Industrial informatics (Přemysl Šůcha)
4. Robotics and machine perception (RMP, Tomáš Pajdla)
   - Robotic perception (K. Štěpánová)
   - Intelligent and mobile robotics (L. Přeučil)
   - Applied algebra and geometry (T. Pajdla)
   - Intelligent machine perception (J.Šivic)
   - Autonomous robots (systems) serve as the unifying theme in the RMP department.
5. Industrial production and automation (Michael Valášek)
6. Cognitive systems and neurosciences (Lenka Vysloužilová)
7. Biomedical engineering and assistive technologies (Olga Štěpánková)

Testbed for Industry 4.0



ROP research interests

Research interests (ROP)

- Industrial robotics
  - Computer vision and perception, geometry
  - Autonomous robotics, mobile robots
  - Perception for self-driving cars
  - Force/torque compliant robots, safety
  - Reactive robot control



- Cognitive robotics and Human ↔ robot collaboration
  - Feedback from sensors/human -> interactive learning
  - Learning from demonstrations and videos
  - Learning from linguistic instructions and gestures
  - Interactive learning
  - Self-supervised learning
  - Knowledge representation
  - Safety
  - Multimodal integration
  - Humanoid robotics







- Benchmarks and datasets, visualisa

Artificial intelligence methods

- Computer vision methods
- Imitation learning
- Machine learning (RL, NN, Bayesian models, generative models), creating statistical models for classification and prediction
- Knowledge graphs - e.g., Ontology-based knowledge representation

MAIN EQUIPMENT:

- Panda Emika collaborative 7DoF robot
- KUKA 7DoF collaborative robot
- HTC Vive Pro with eytracker + Hi5 glove
- Leap Motion with tactile feedback
- Several Realsense cameras



DEVELOPMENT ENVIRONMENT:

- ROS2 integration
- Coppelia simulator, PyBullet, iGibson
- Ontology database

Excellency at CIIRC CTU in AI

- ELLIS - the European Laboratory for Learning and Intelligent Systems - is a pan-European AI network of excellence which focuses on fundamental science, technical innovation and societal impact. Founded in 2018, ELLIS builds upon machine learning as the driver for modern AI
- Claire, Confederation of Laboratories for Artificial Intelligence in Europe, https://claire-ai.org/
- **Dr. Josef Šivic, doc. Tomáš Pajdla a** Dr. TomáŠ Mikolov from CIIRC scored among the 10 best Czech Scientists in the world wide scientists list Guide2Research v Top Computer Scientists Ranking.

About me

- Msc – Condensed matter physics, MFF UK
- PhD - Artificial Intelligence and Biocybernetics from CTU, FEE, Dept. of Cybernetics, Topic of Dissertation thesis: Hierarchical probabilistic model of language acquisition
- Researcher at CIIRC CTU in Prague (RMP group), Head of ROP group

Research interests

- Probabilistic models of cognition
- Multimodal learning
- Unsupervised learning
- Imitation learning with a support of language and gestures
- Task representation (e.g., assembly via behavior trees)
- HRI communication and dialogues
- Robot self-calibration



Imitrob team

CIIRC CTU in Prague

http://imitrob.ciirc.cvut.cz

Github: https://github.com/imitrob



ROP group

CIIRC CTU in Prague

http://rop.ciirc.cvut.cz



ELLIOT: European Large Open Multi-Modal Foundation Models For Robust Generalization On Arbitrary Data Streams

ELLIOT is a four-year Horizon European project building open, trustworthy Multimodal Generalist Foundation Models (MGFMs) for real-world use.

- Robotic perception usecase (lead CTU)
  - Adding new modalities
  - Automatic trajectory generation from 3D data and demonstrations
    - The system **learns coating motions** from human demonstrations
    - It uses 3D object models to **automatically generate pre-coating paths that respect the geometry of real parts**.
  - Adaptation to object variations and repairs
    - **Trajectories are automatically adjusted** when objects move, change in size, shape or segment count.
    - Vision models identify **insufficient or defective coating regions**
  - Intuitive, multimodal interface and continuous safety
    - Operators can **specify coating requirements** using natural language and gestures
  - Usecase provider: Robotwin, s.r.o.





PIs (CTU): **J.Sivic, K.Stepanova, V.Petrik**
ROP team: M.David, L.Wagner, M.Majsner

Project: EU ELLIOT (2025-2029)

https://elliot-ai.eu/project-overview

Responsible person: karla.stepanova@cvut.cz

PersonalRobot: Customizing Robots via Multimodal Interactive Human-Robot Dialogues (GACR Junior STAR, 2026-31)

The goal is a robot that remembers past instructions, reasons under uncertainty, and *knows when it doesn't know enough while taking into account the context of the human (attention, mental load, preferences) and environment (available tools, etc.)*.

- Four main subtasks
  - Updating task knowledge from multimodal partial observation
  - Context - modeling human and environment
  - Identifying knowledge gaps in a robot memory system
  - Interactive multimodal dialogue for task knowledge refinement



Team: **K.Stepanova**, R.Skoviera, P.Krsek,P.Vanc, M.Vavrecka

Project: GACR Junior STAR (2026-2031)

Responsible person: karla.stepanova@cvut.cz

Proč již nestačí roboti na jednu úlohu?



specializovaný průmyslový robot

Proč již nestačí roboti na jednu úlohu?



specializovaný průmyslový robot

Automatizace malovýroby
proměnlivý svět,
nové úlohy,
Personalizace
Sdílený prostor s lidmi





Proč již nestačí roboti na jednu úlohu?



specializovaný průmyslový robot

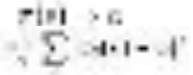
Automatizace malovýroby
proměnlivý svět,
nové úlohy,
Personalizace
Sdílený prostor s lidmi





Robot učící se od člověka, Robotwin s.r.o.





Calibration
Object pose detection

Robot spolupracující s člověkem





Flexibilní humanoid, Atlas Boston dynamics

https://www.youtube.com/watch?v=QgevJRna_hk

Jak toho dosáhnout

1. Více dat a větší modely
2. Málo dat a jednodušší modely

Vyřeší vše více dat?



Z teleoperací…





Inside China's 'robot school' where humanoid machines are learning everyday tasks

https://mobile-aloha.github.io/
Team at Stanford University

Fu, Zipeng, Tony Z. Zhao, and Chelsea Finn. "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation." *Conf. on Robot Learning* (2024).

A dalších teleoperací…



M. Beetz, University of Bremen, Germany



https://vrb.ease-crc.org/explore-labs/neem-lab/

Nebo z videí?



Miech, Antoine, et al. "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

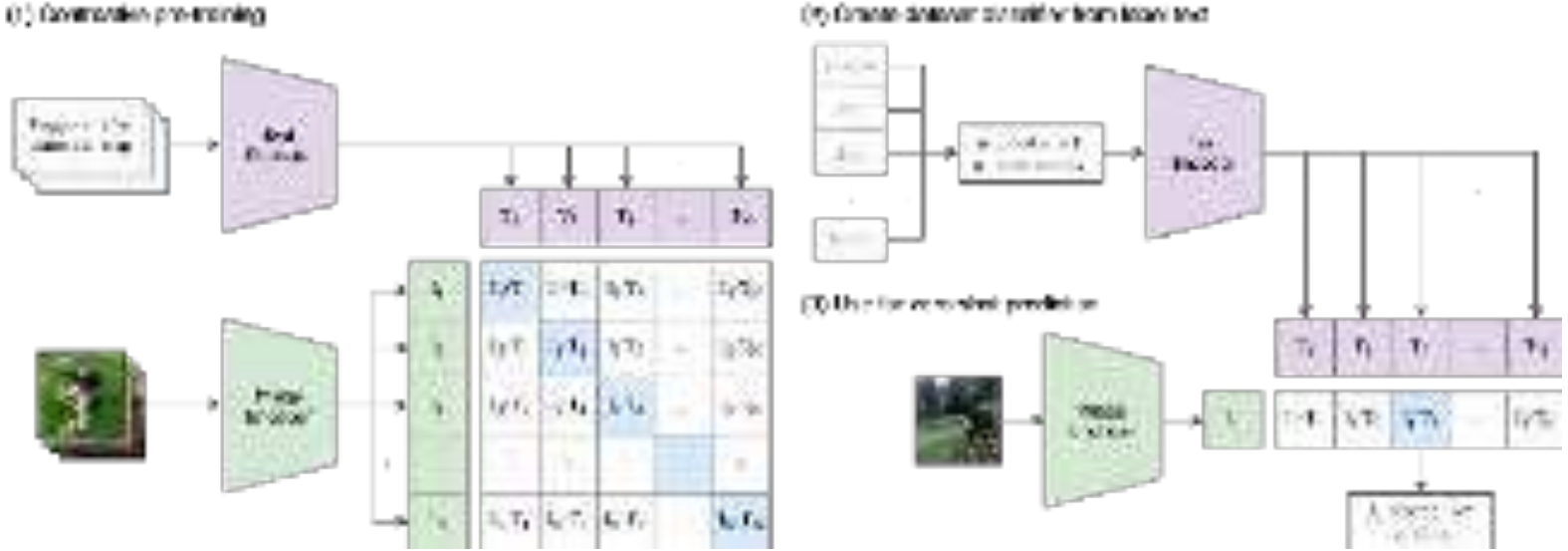
France + CR



Gavryushin, Alexey, et al. "Maple: Encoding dexterous robotic manipulation priors learned from egocentric videos." *arXiv preprint arXiv:2504.06084* (2025).

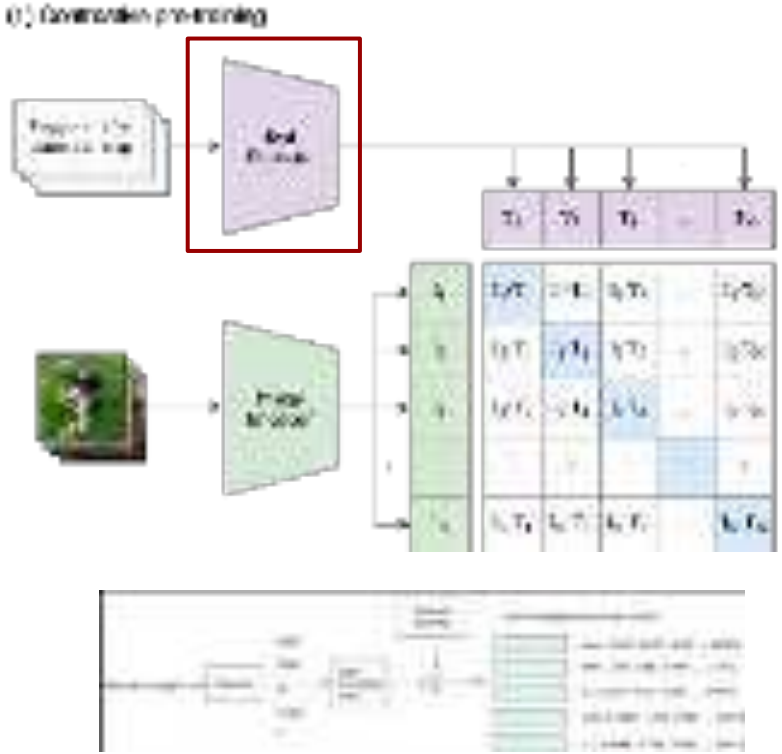
ETH Zurich

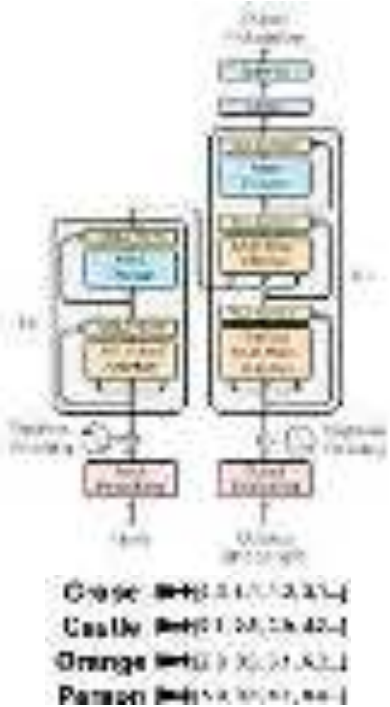
VLM - CLIP-based models

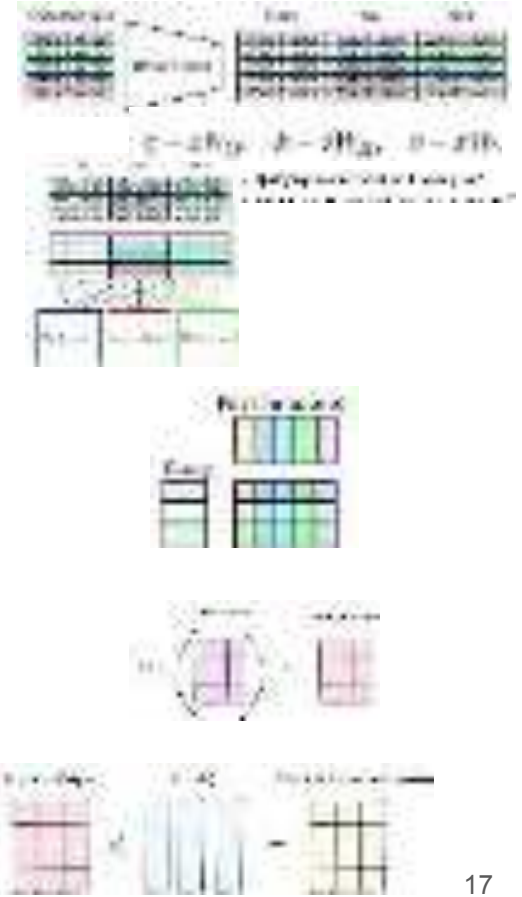


Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021.

VLM - CLIP-based models

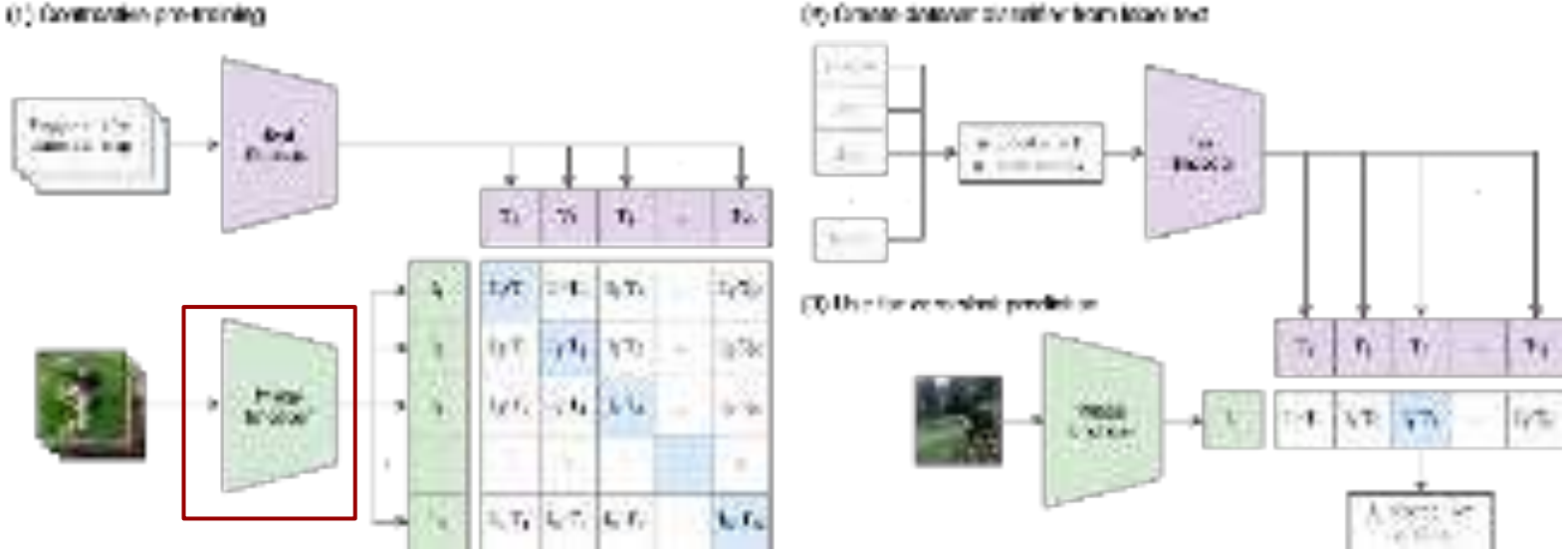






https://towardsdatascience.com/clip-intuitively-and-exhaustively-explained-1d02c07dbf40/

VLM - CLIP-based models

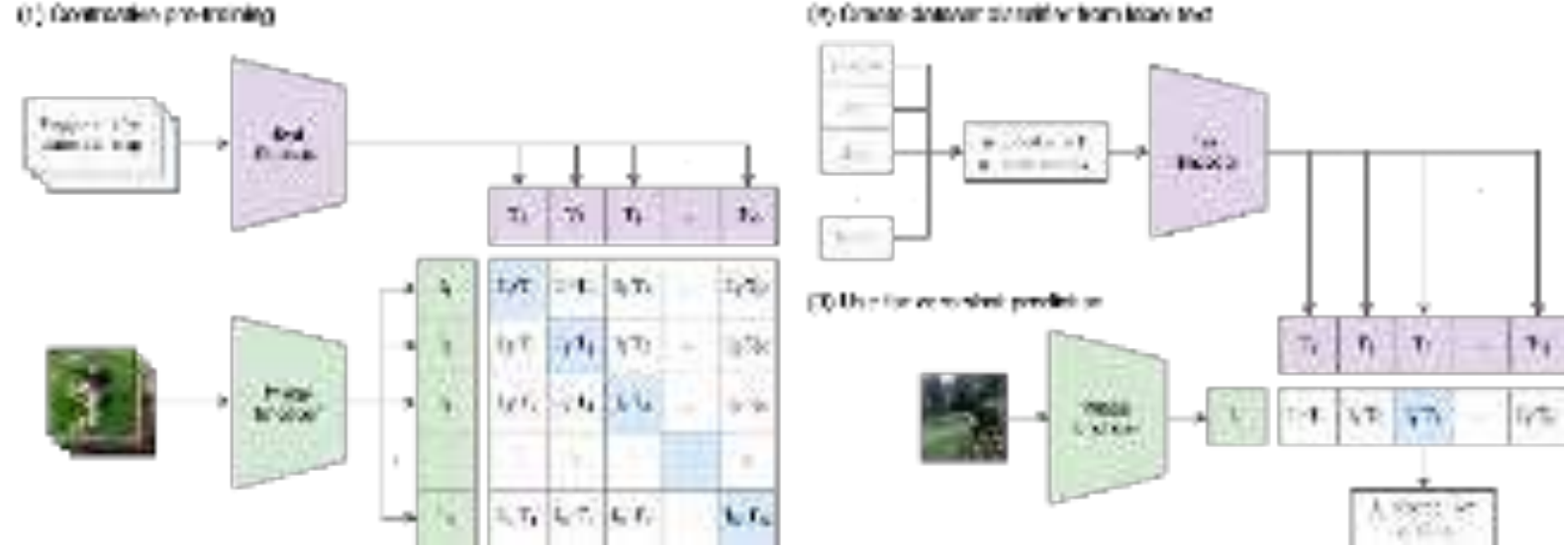






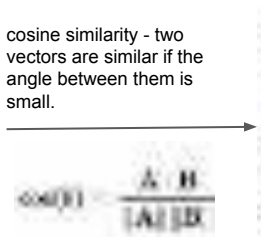
https://towardsdatascience.com/clip-intuitively-and-exhaustively-explained-1d02c07dbf40/

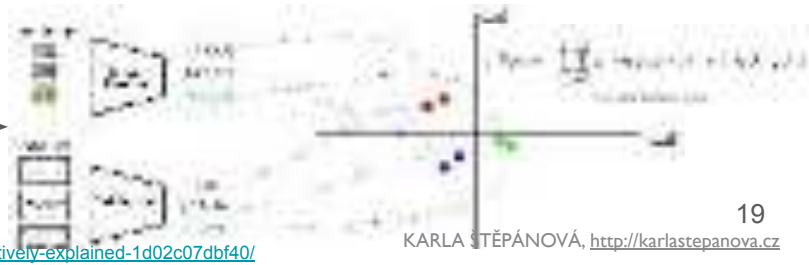
VLM - CLIP-based models





cosine similarity - two vectors are similar if the angle between them is small.

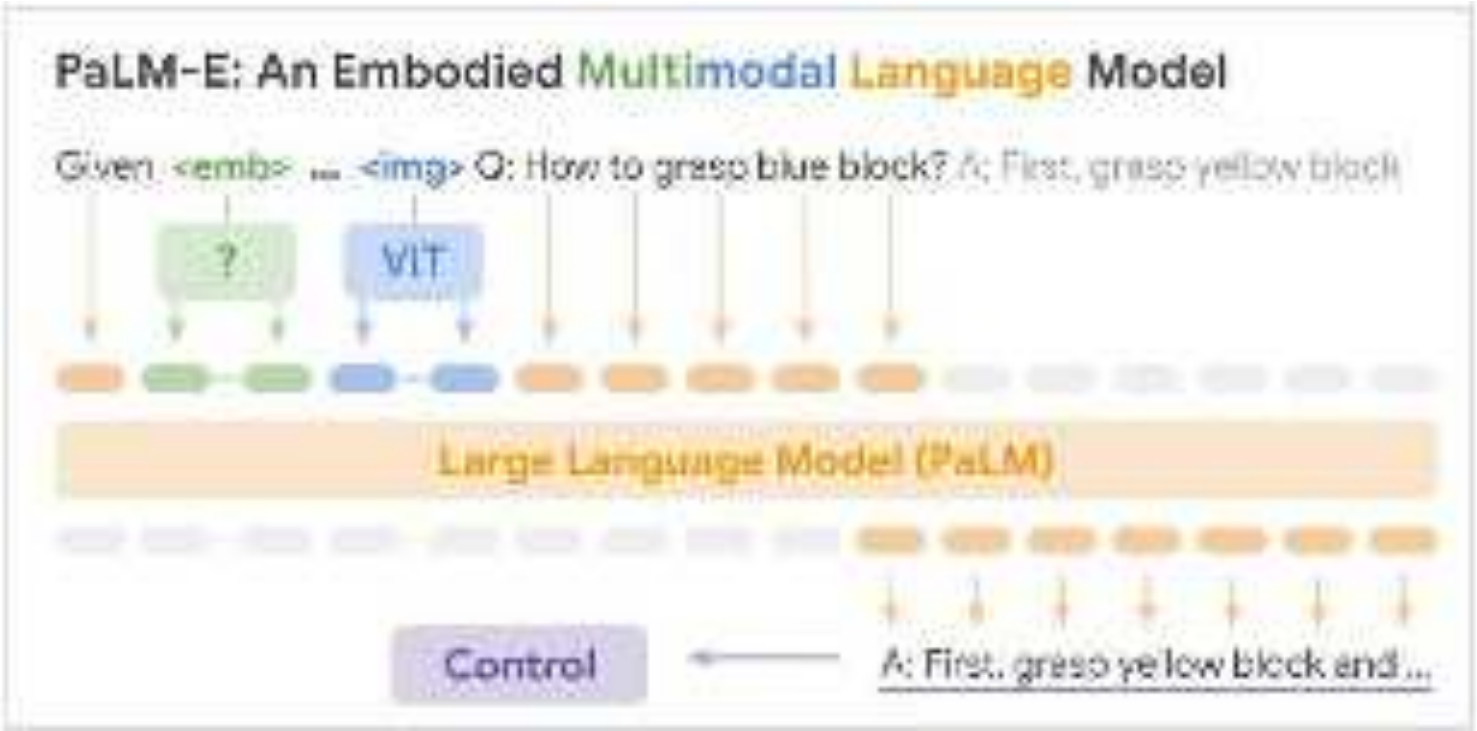




https://towardsdatascience.com/clip-intuitively-and-exhaustively-explained-1d02c07dbf40/


Contrastive Loss
Generative Loss
Image Feat.
Text Feat.
Cross Attention
"A white cliff with plants on it"
Image
Text
MaMMUT (Ours)
Generative Loss
Image Feat.
Text Feat.
Image
Text
Encoder-Decoder
Generative Loss
Contrastive Loss
Image Feat.
Text Feat.
"A white cliff with plants on it"
Image
Text
Contrastive Alignment and Captioning

PALM-E

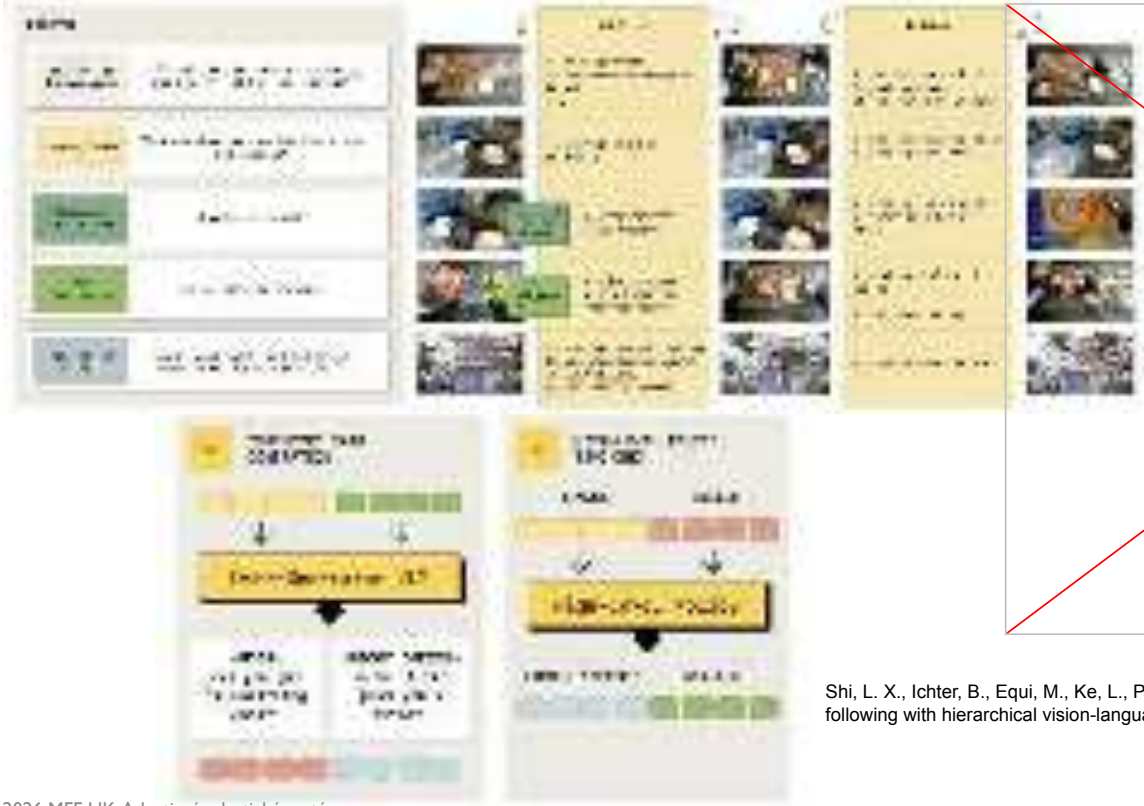

PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp ye low block and ...

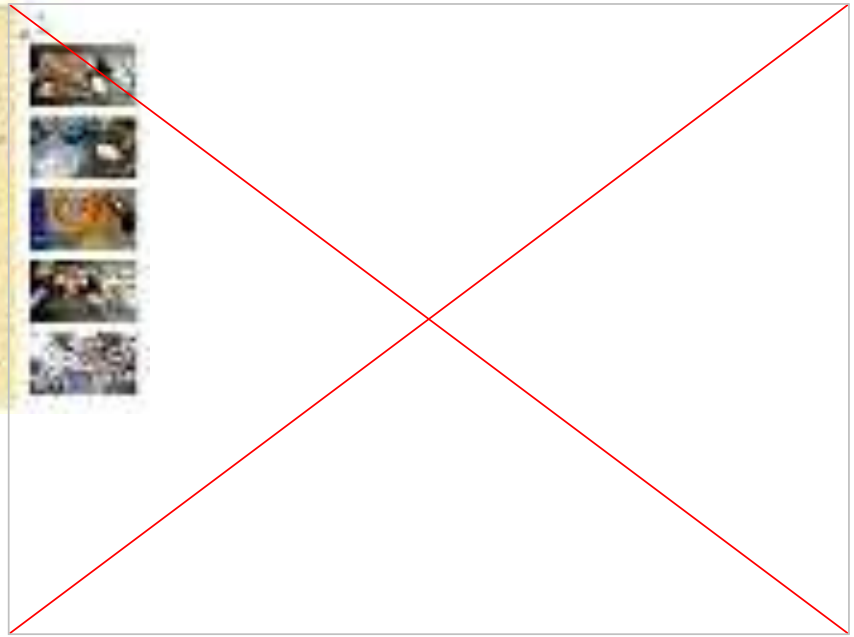
VLA, pi0


OpenVLA
Action De-Tokenizer
Llama 2 7B
Llama Tokenizer
DinoV2
SigLIP



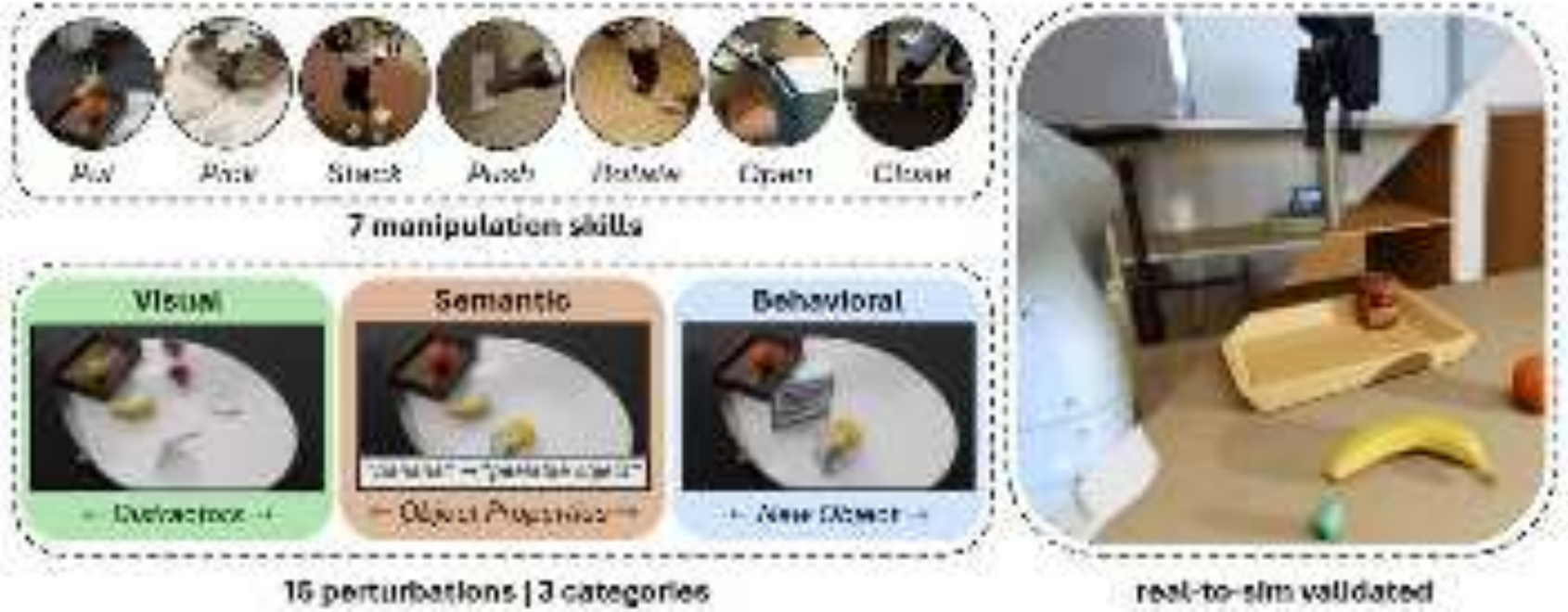
HiRobot





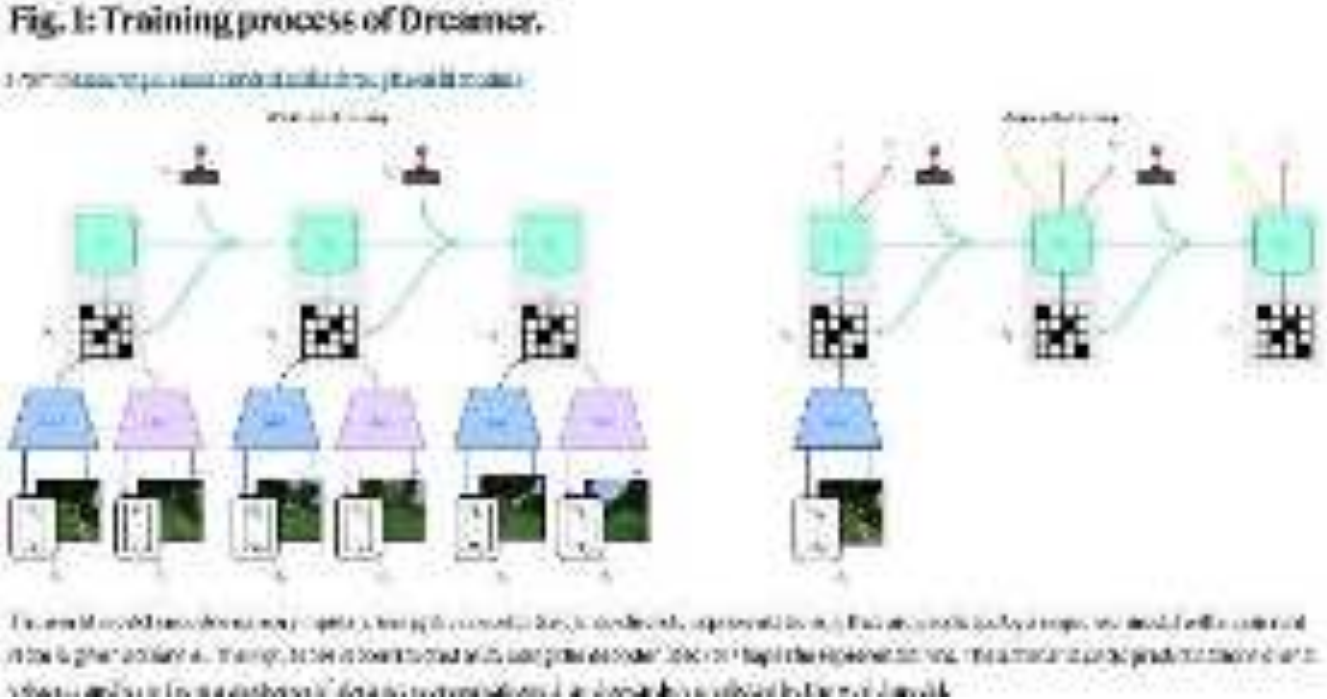
Shi, L. X., Ichter, B., Equi, M., Ke, L., Pertsch, K., Vuong, Q., ... & Finn, C. (2025). Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*.


Generalist
Policy Pool
Distributed Network of Evaluators
RoboArena
Policy
Ranking
Policy
Score



Sedlacek, Martin, et al. "REALM: A Real-to-Sim Validated Benchmark for Generalization in Robotic Manipulation." *arXiv preprint arXiv:2512.19562* (2025).

World models as a simulator for robotics



https://marble.worldlabs.ai/

Intuitive physics

COSMOS POLICY: FINE-TUNING VIDEO MODELS FOR VISUOMOTOR CONTROL AND PLANNING

Moo Jin Kim[1,2] Yihuai Gao[1,3] Tsung-Yi Lin[1] Yen-Chen Lin[1] Yunhao Ge[1] Grace Lam[1]
Percy Liang[2] Shuran Song[1,3] Ming-Yu Liu[1] Chelsea Finn[2] Jinwei Gu[1]
[1]NVIDIA [2]Stanford University

https://research.nvidia.com/labs/dir/cosmos-policy/

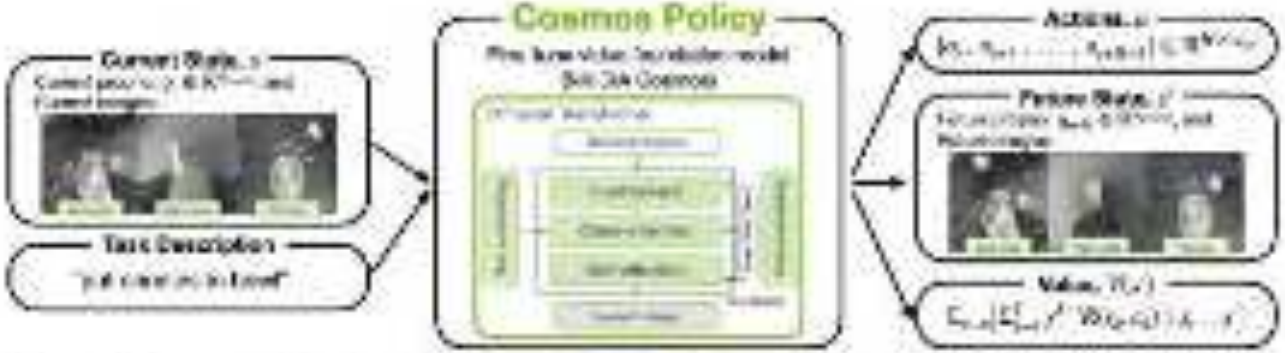
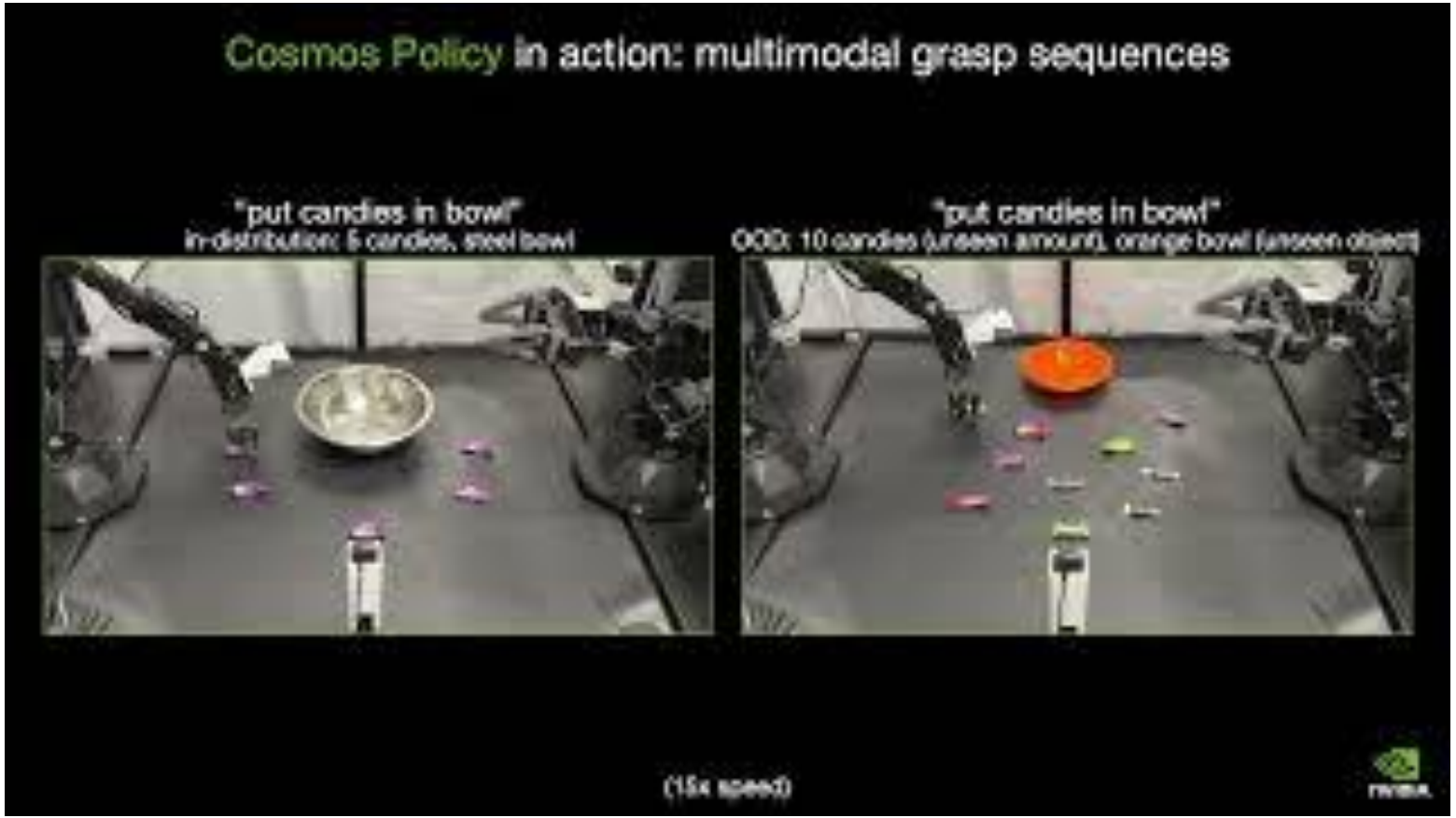


Figure 1: We present Cosmos Policy, a state-of-the-art robot policy fine-tuned from the NVIDIA Cosmos-Predict2-2B video foundation model. Cosmos Policy handles multimodal inputs and multi-view camera images and predicts (1) robot actions chunk, (2) future state (represented by robot proprioception and image observations), and (3) value (expected rewards-to-go at the future state). No architectural changes are made to the base video model, and all modalities are jointly modeled through the natural video denoising objective.


Cosmos Policy in action: multimodal grasp sequences
"put candies in bowl"
in-distribution: 5 candies, steel bowl
"put candies in bowl"
OOD: 10 candies (unseen amount), orange bowl (unseen object)
(15x speed)


Action
Text
Text DeTokenizer
Action Model
Text Tokenizer
Image
Text
(a) Action Model
(e.g., OpenVLA)
Image Understanding
Image Generation
Action Understanding
Action Generation
Image
Text
Image DeTokenizer
Text DeTokenizer
World Model
Image Tokenizer
Text Tokenizer
Action Tokenizer
Image
Text
Action
(b) World Model
(e.g., iVideoGPT)
Image Understanding
Image Generation
Action Understanding
Action Generation
Image
Text
Action
Image DeTokenizer
Text DeTokenizer
Action DeTokenizer
Action World Model
Image Tokenizer
Text Tokenizer
Action Tokenizer
Image
Text
Action
(c) Action World Model
(e.g., WorldVLA)
Image Understanding
Image Generation
Action Understanding
Action Generation

What if we do not have the data?
How can we infer so much from so little?



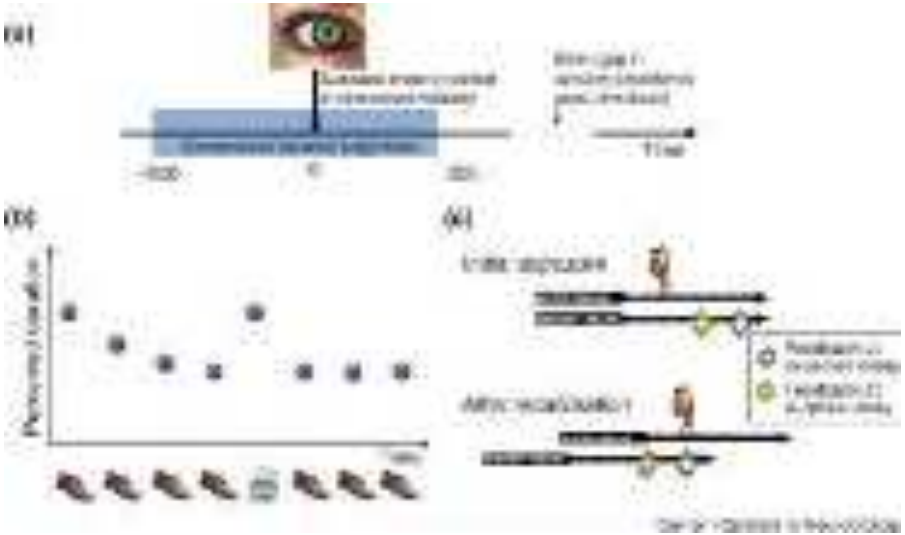
How can we infer so much from so little?



How can we infer so much from so little?

Consciousness and attention
→ Only simulation of reality

- Necessity to create explainable world, usage of apriori information
- Able to integrate noisy input information – prior knowledge gets higher importance



Eagleman DM. Human time perception and its illusions. Curr Opin Neurobiol. 2008 Apr;18(2):131-6. doi: 10.1016/j.conb.2008.06.002. Epub 2008 Aug 8. PMID: 18639634; PMCID: PMC2866156.

How can we infer so much from so little?

Consciousness and attention
→ Only simulation of reality

- Necessity to create explainable world, usage of apriori information



How can we infer so much from so little?

Consciousness and attention
→ Only simulation of reality

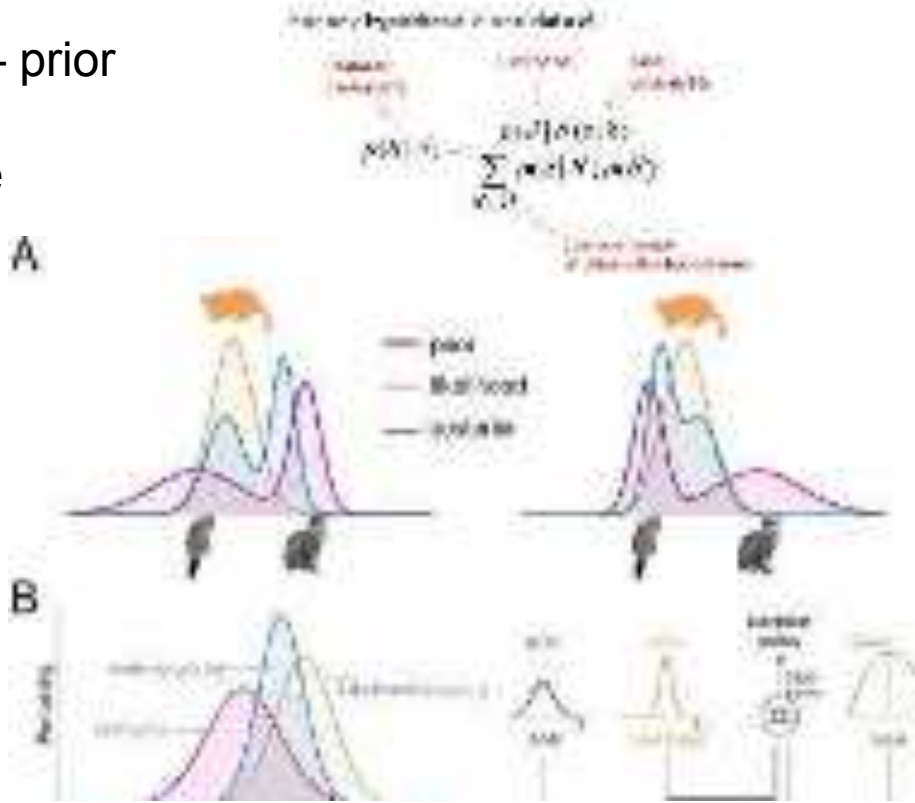
- Necessity to create explainable world, usage of apriori information
- Able to integrate noisy input information – prior knowledge gets higher importance
- How can we infer so much from so little evidence?
    - Deduction? Intuition?
    - Concepts from examples



How can we infer so much from so little?

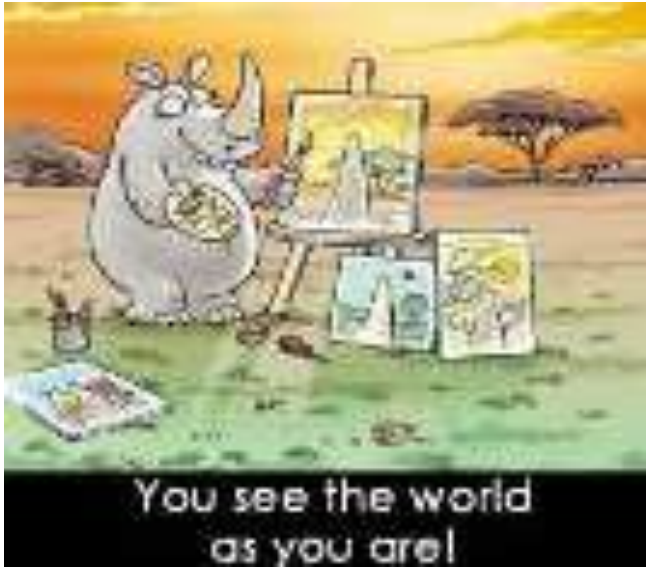
- Able to integrate noisy input information – prior knowledge gets higher importance
  - How can we infer so much from so little evidence?
  - Deduction? Intuition?
  - Concepts from examples



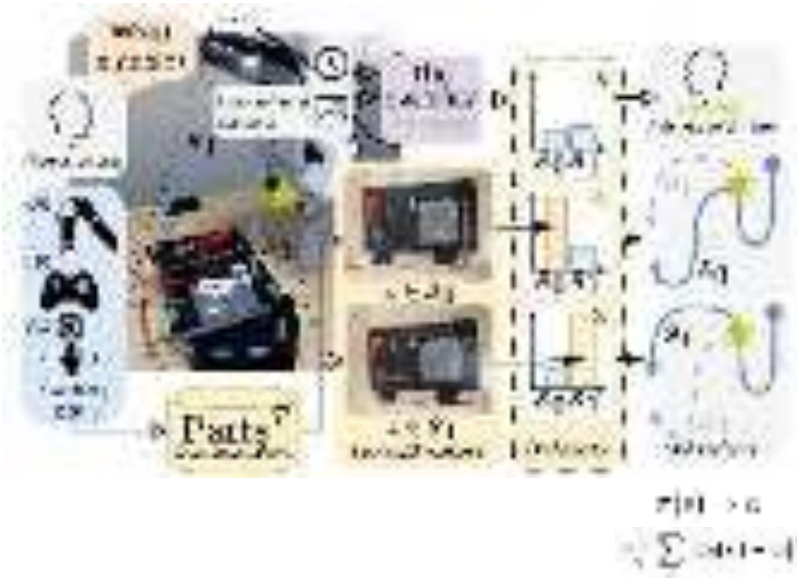


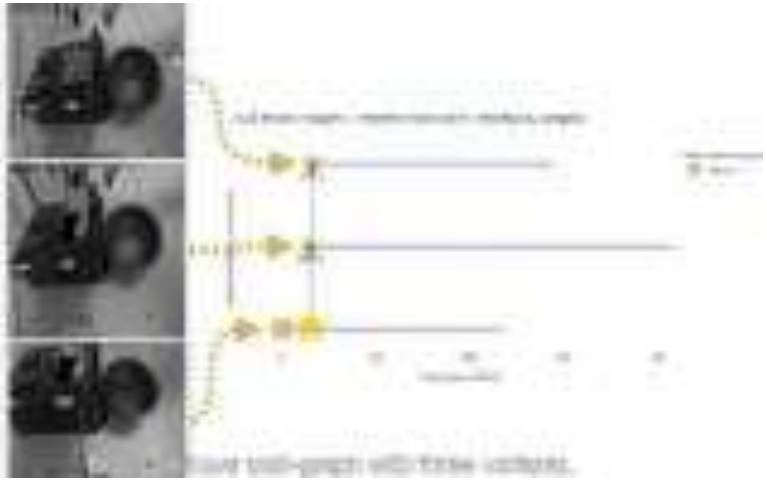
How can we infer so much from so little?

- Able to integrate noisy input information – prior knowledge gets higher importance
- How can we infer so much from so little evidence?
  - Deduction? Intuition?
  - Concepts from examples
  - Human learning via communication
    - Unsupervised/weakly-supervised
    - Multimodal integration (gestures, eye-contact,...)
    - Ability to correct our actions, include feedback
    - Effect of context – meaning of gestures/words
    - Grounding of knowledge
    - Adaptive responses/behavior



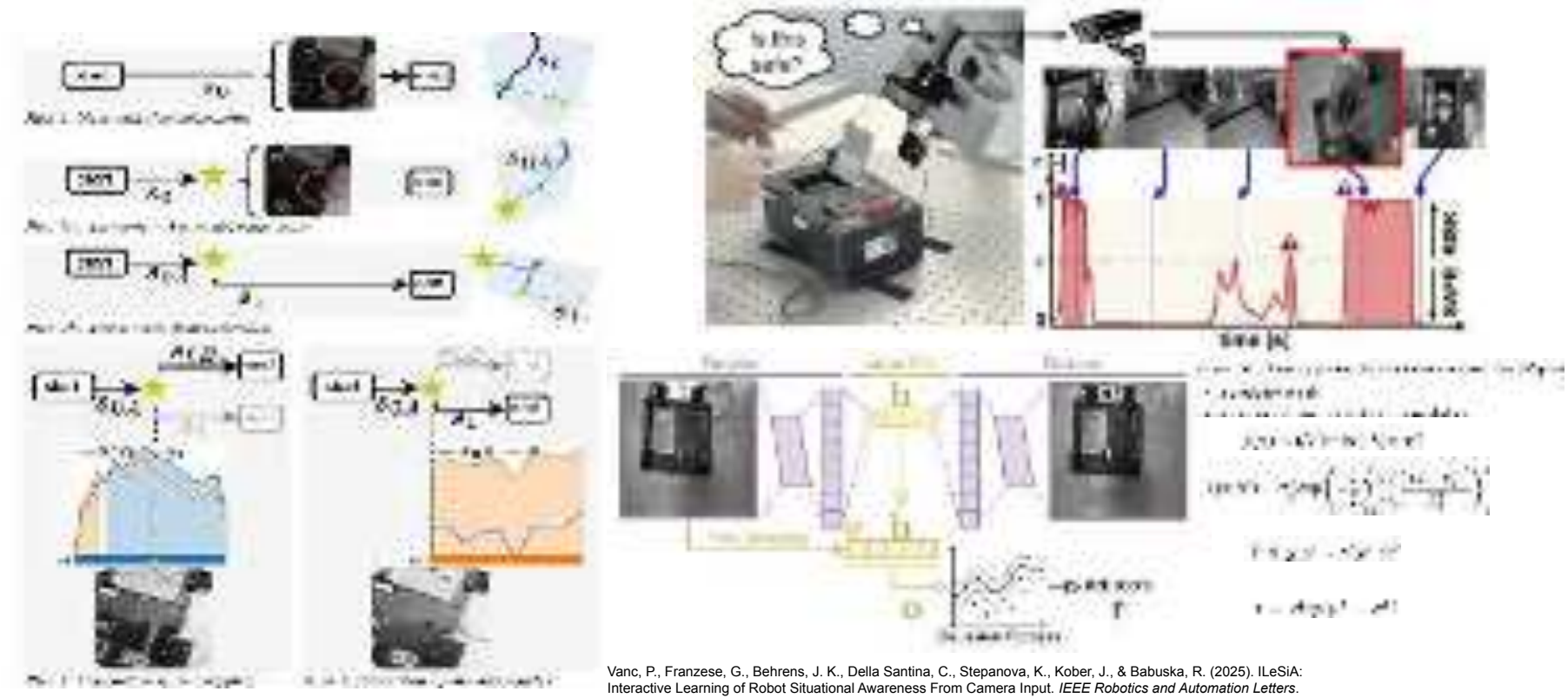
How to learn from few data?
Interactive robot skill programming





Vanc, P., Behrens, J. K., Hlaváč, V., & Stepanova, K. (2026). See and Switch: Vision-Based Branching for Interactive Robot-Skill Programming. *arXiv preprint arXiv:2603.08057*.

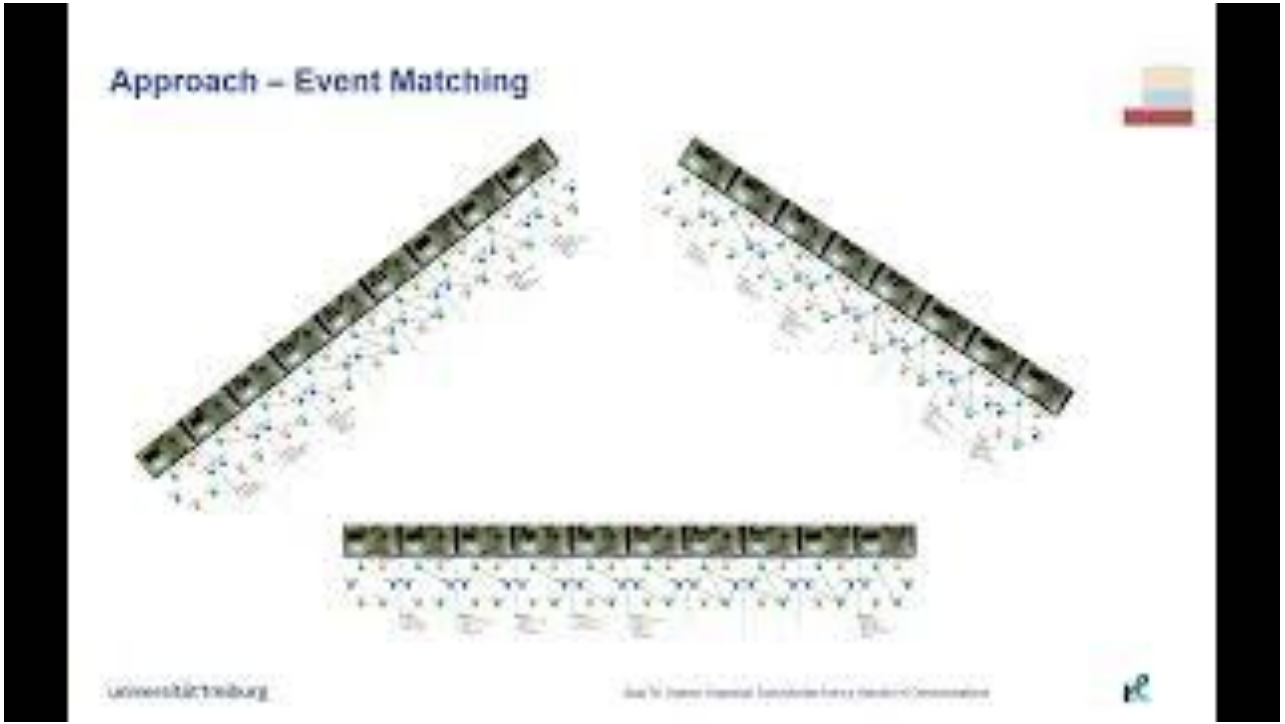
Interactive robot skill programming



Vanc, P., Franzese, G., Behrens, J. K., Della Santina, C., Stepanova, K., Kober, J., & Babuska, R. (2025). ILeSiA: Interactive Learning of Robot Situational Awareness From Camera Input. *IEEE Robotics and Automation Letters*.

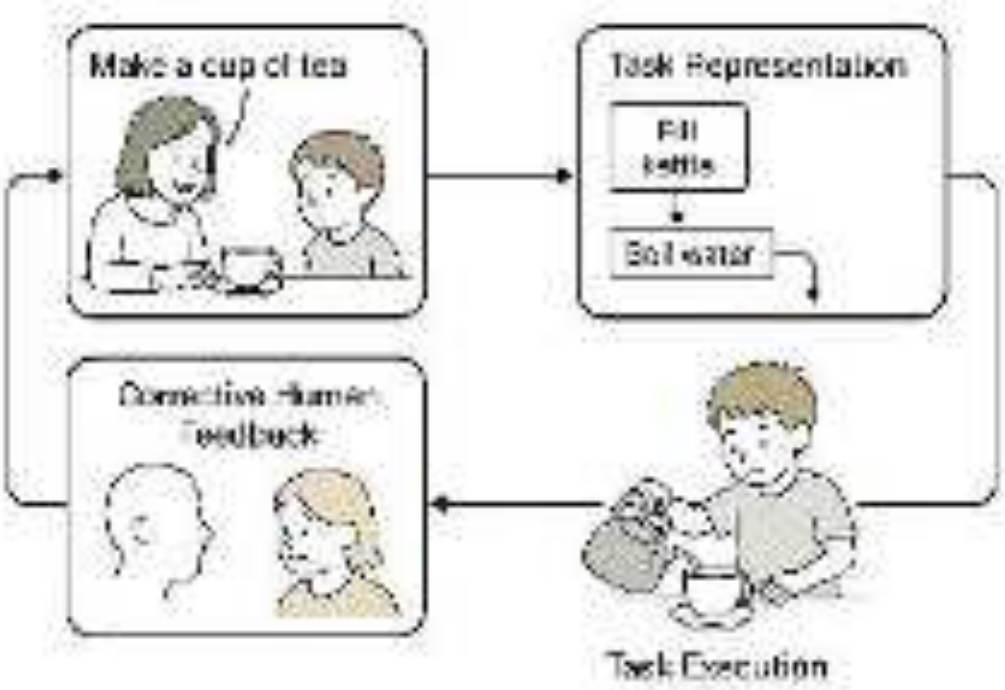
Vanc, P., Behrens, J. K., Hlaváč, V., & Stepanova, K. (2026). See and Switch: Vision-Based Branching for Interactive Robot-Skill Programming. *arXiv preprint arXiv:2603.08057*.

How to learn from few data?



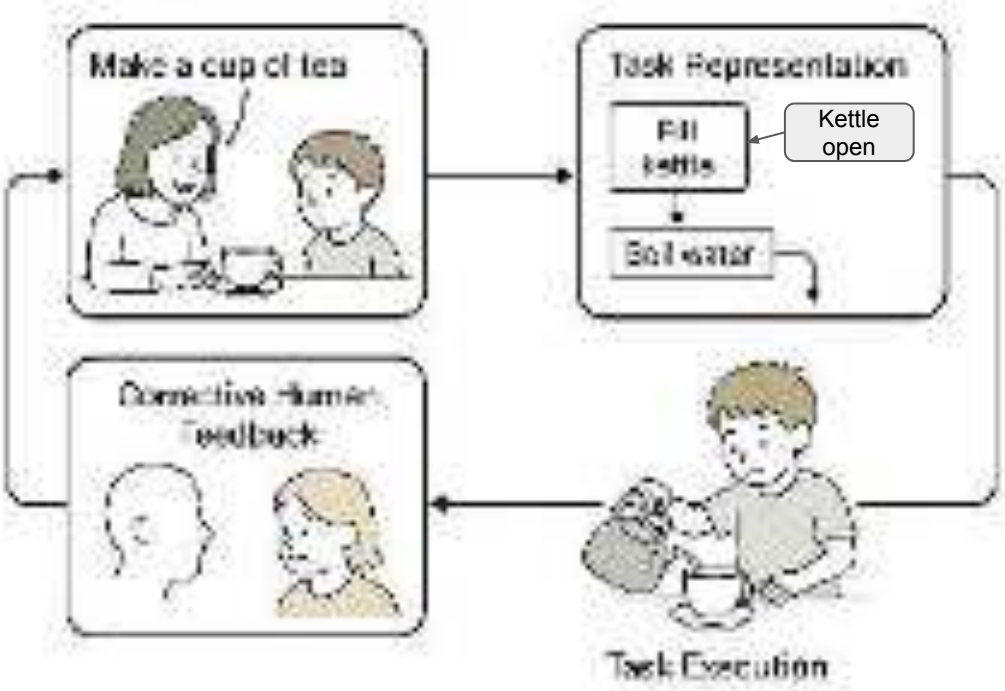
Röfer, Adrian, Nick Heppert, and Abhinav Valada. "SparTa: Sparse Graphical Task Models from a Handful of Demonstrations." *arXiv preprint arXiv:2602.16911* (2026).

Co potřebujeme pro naučení robota úlohy?



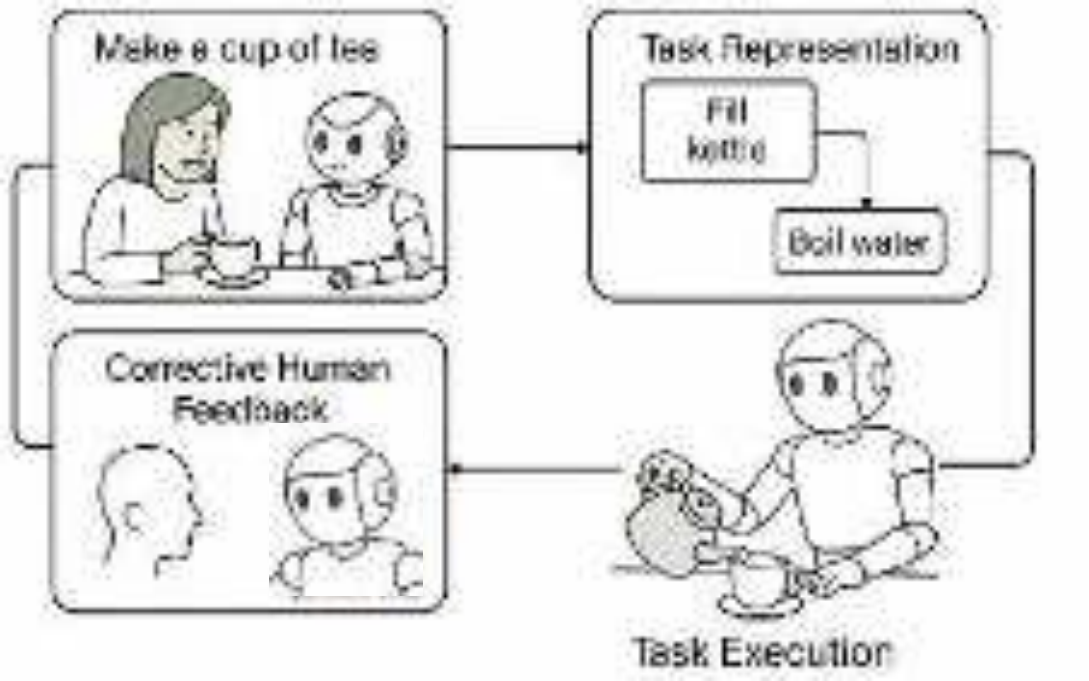
True intelligence isn't knowing everything—it's knowing how to learn from others.

Co potřebujeme pro naučení robota úlohy?



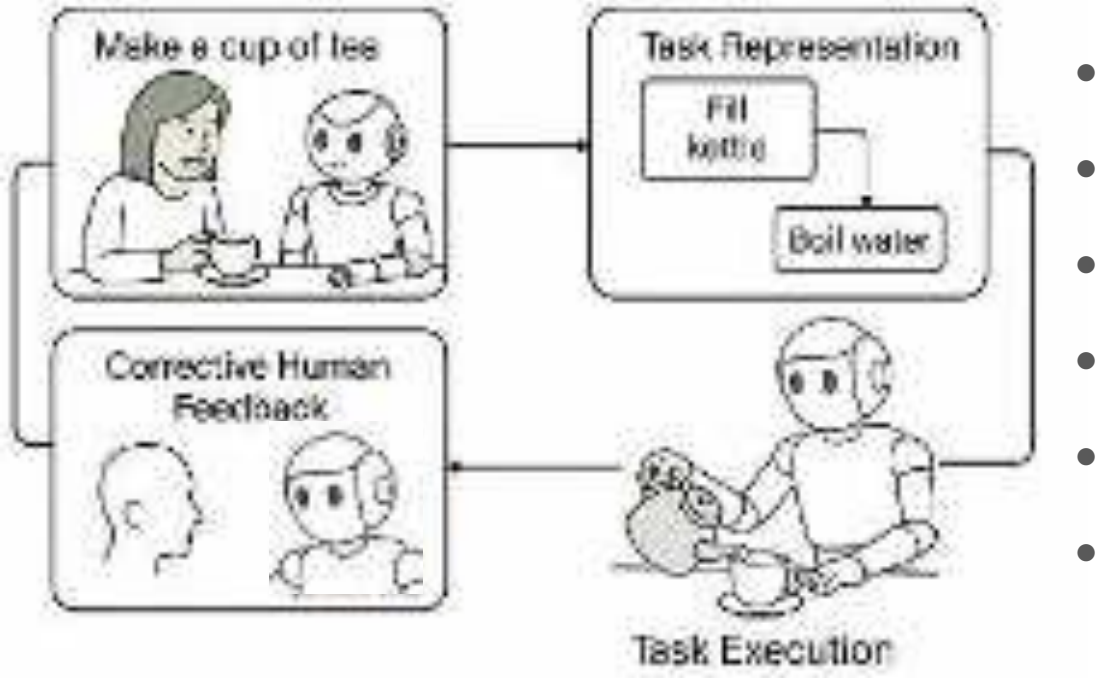
True intelligence isn't knowing everything—it's knowing how to learn from others.

Co potřebujeme pro naučení robota úlohy?



True intelligence isn't knowing everything—it's knowing how to learn from others.

Co potřebujeme pro naučení robota úlohy?



- **Rozumět** tomu, co mu člověk říká
- **Chápat pokyny, demonstrace** i **omezení** (slova, gesta, situaci)
- Vytvořit si představu o tom, **co je cílem úlohy**
- **Vykonávat** úlohu a reagovat na změny v okolí
- Umět si **vyžádat zpětnou vazbu** a poučit se z chyb
- **Přenést naučené zkušenosti** na nové úlohy, prostředí nebo uživatele

True intelligence isn't knowing everything—it's knowing how to learn from others.

Co potřebujeme pro naučení se úlohy

- Schopnost naučit se danou úlohu a vykonat ji přesně a opakovaně v různých podmínkách
- Porozumění úloze jako celku
- Chápání prostředí a jeho změn
- Schopnost odhadnout záměr člověka
- Učení z demonstrací a oprav/zpětné vazby člověka
- Schopnost robota aktivně se doptávat, získávat nové znalosti
- Schopnost robota opravit úlohu
- Přizpůsobení se různým uživatelům

1. Pochopení významu slov

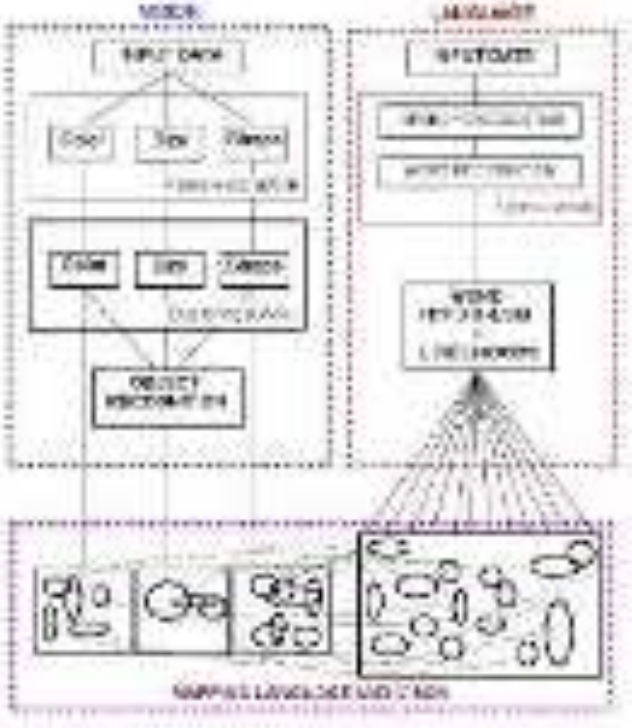
How do symbols (words) acquire meaning from perceptual data?

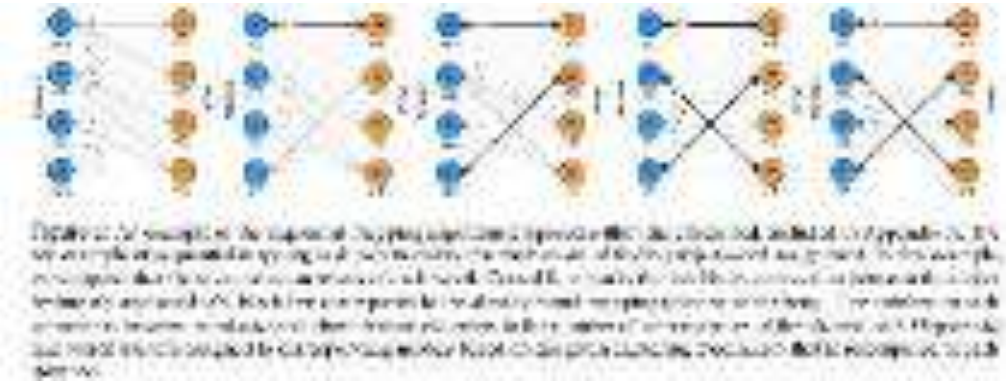
“Gavagai kulabai”

- Only language (tone)
- Language + vision
- Language + gesture/eye-gaze
- Language + demonstration
- Proprioception, tactile feedback, etc.

1. Pochopení významu slov

Word-to-object mapping



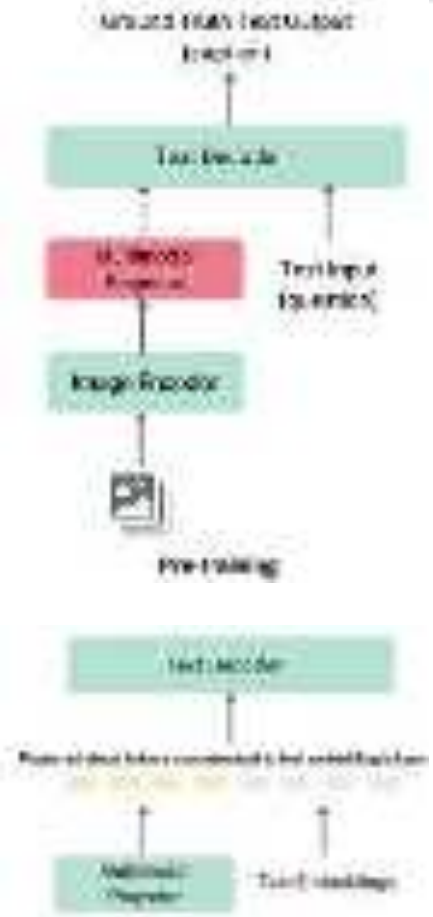


Štěpánová K., Frederico B. Klein, Angelo Cangelosi, Vavrečka M. “Mapping language to vision in a real-world robotic scenario.” IEEE Transactions on Cognitive and Developmental Systems (2018). Paper, video

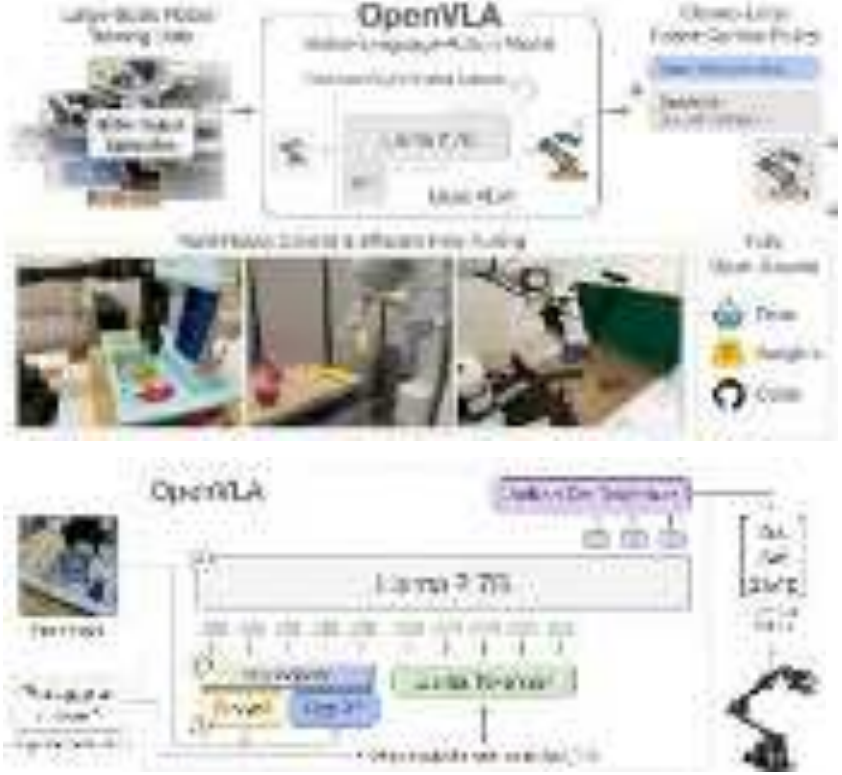
1. Pochopení významu slov

Word-to-action mapping





1. Pochopení významu slov





1. Pochopení významu slov

Word-to-action mapping





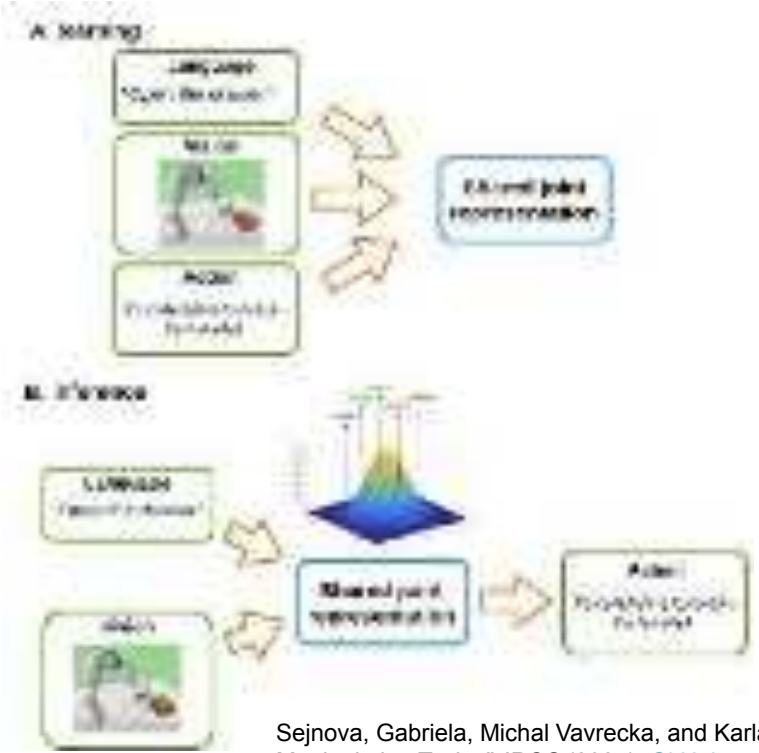
1. Pochopení významu slov





1. Pochopení významu slov

Word-to-action mapping

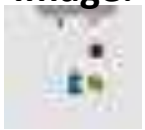


- Alternative to “VLA” models
- Apply mVAE to robotics data (generate trajectories based on language instructions)
- Cross-generation

Impact: Enabling to learn novel tasks. Robot can show what it understood under the task.

Instruction:
“move left the apple”

Image:



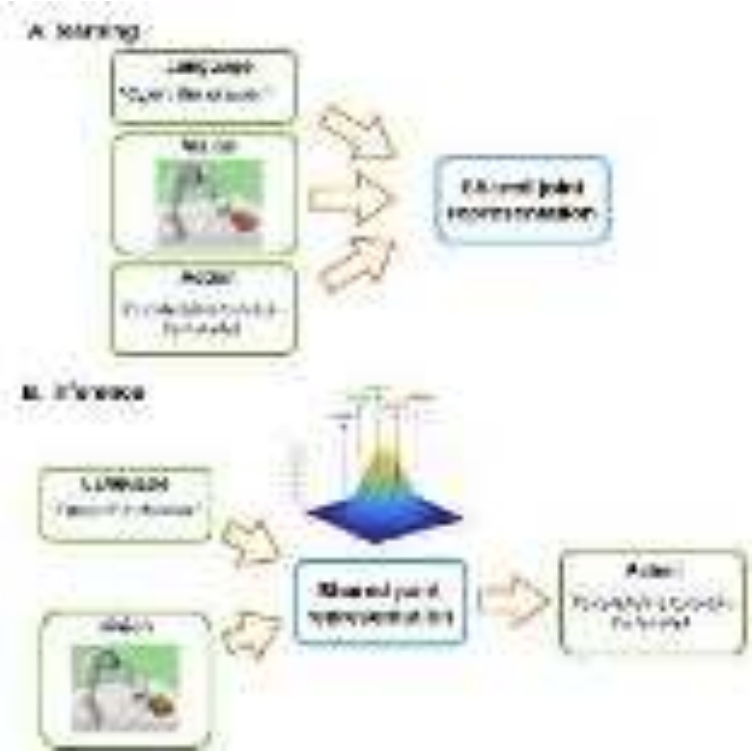
Generated trajectory



Sejnova, Gabriela, Michal Vavrecka, and Karla Stepanova. "Bridging Language, Vision and Action: Multimodal VAEs in Robotic Manipulation Tasks." *IROS* (2024), GitHub, paper on ArXiv. MVAE toolkit: GitHub.

1. Pochopení významu slov

Word-to-action mapping

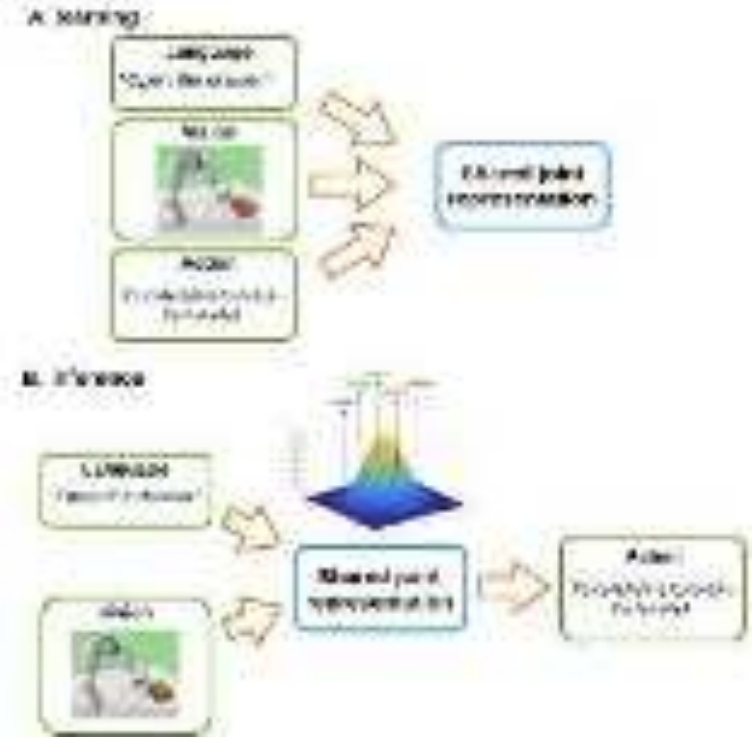




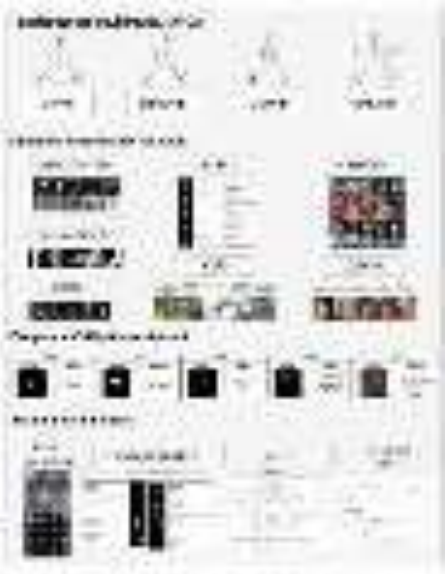
1. Pochopení významu slov

Word-to-action mapping

$$\mathcal{L}(x) = \mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \| p(z))$$

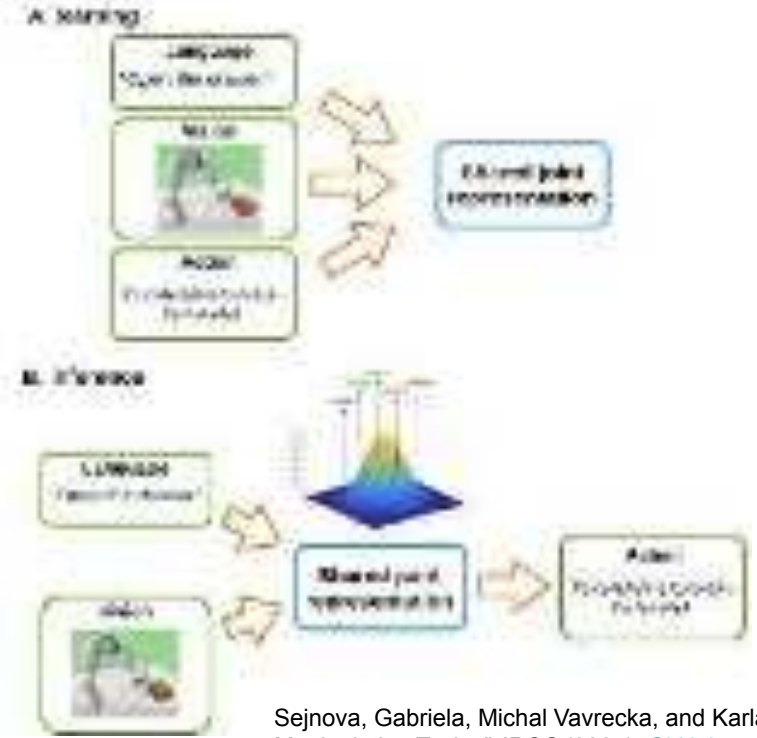






1. Pochopení významu slov

Word-to-action mapping

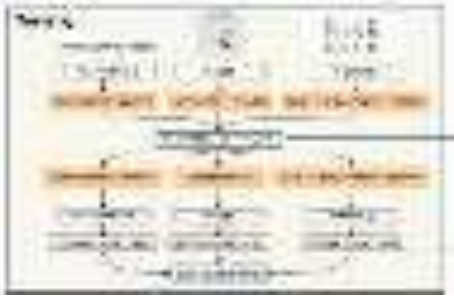


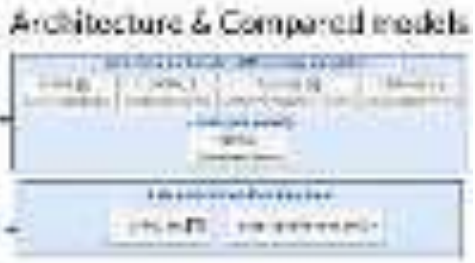
Sejnova, Gabriela, Michal Vavrecka, and Karla Stepanova. "Bridging Language, Vision and Action: Multimodal VAEs in Robotic Manipulation Tasks." *IROS* (2024), GitHub, paper on ArXiv. MVAE toolkit: GitHub.

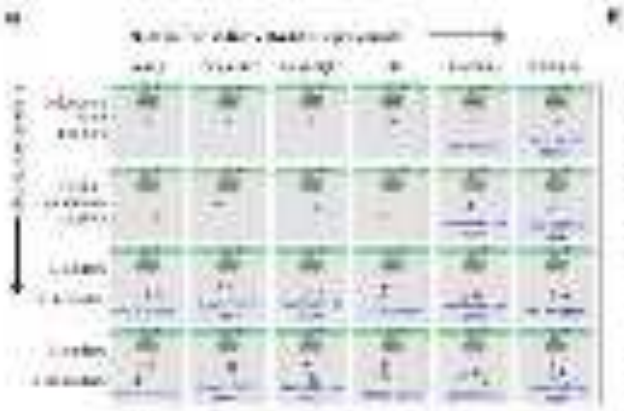
1. Pochopení významu slov

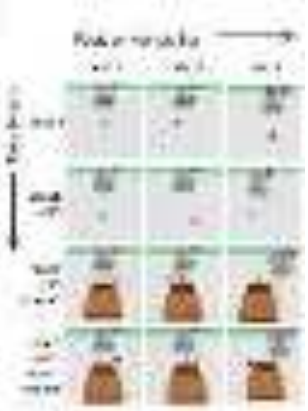
Word-to-action mapping







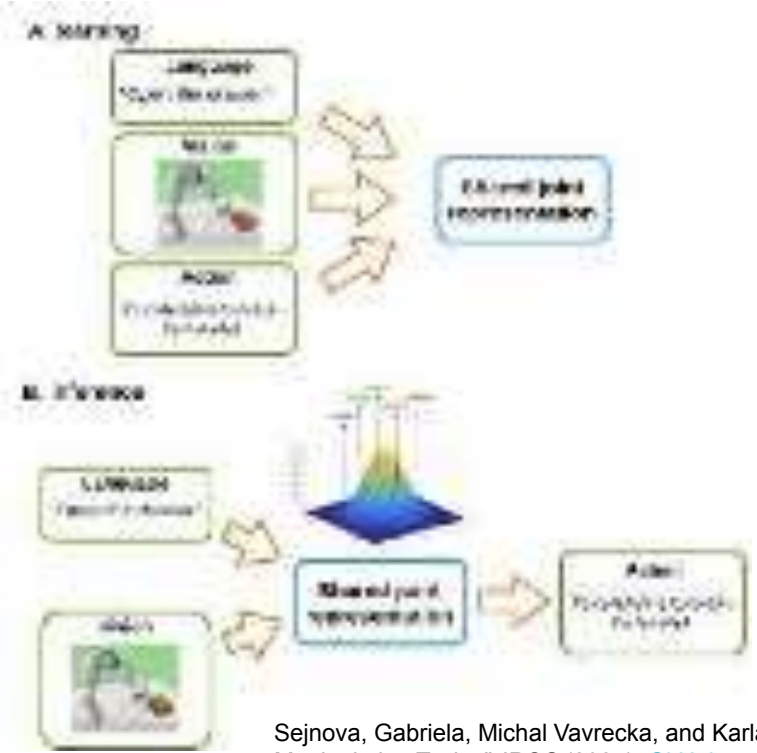


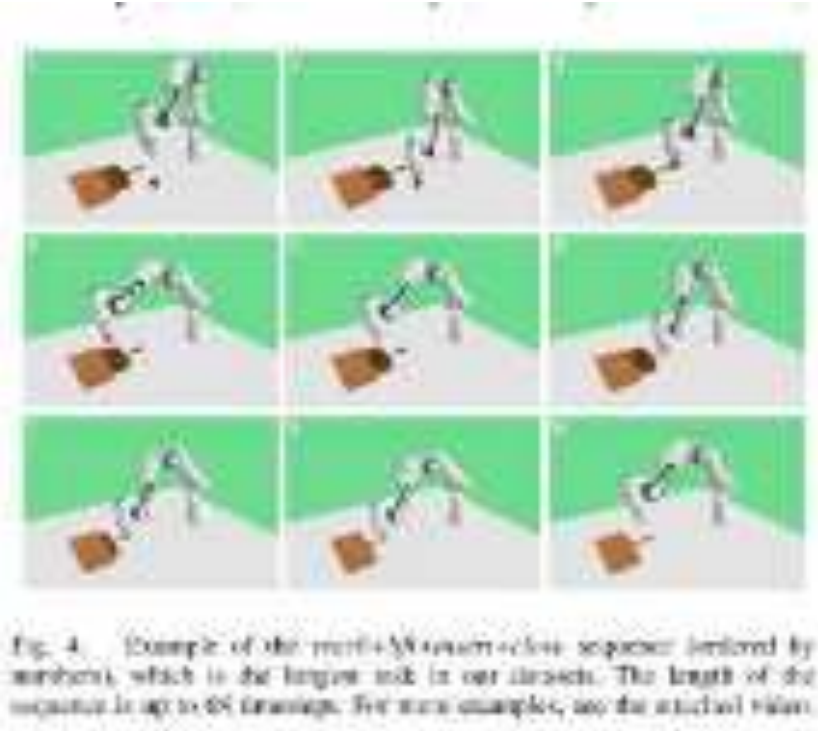


Sejnova, Gabriela, Michal Vavrecka, and Karla Stepanova. "Bridging Language, Vision and Action: Multimodal VAEs in Robotic Manipulation Tasks." *IROS* (2024), GitHub, paper on ArXiv. MVAE toolkit: GitHub.

1. Pochopení významu slov

Word-to-sequence of actions





Sejnova, Gabriela, Michal Vavrecka, and Karla Stepanova. "Bridging Language, Vision and Action: Multimodal VAEs in Robotic Manipulation Tasks." *IROS* (2024), GitHub, paper on ArXiv. MVAE toolkit: GitHub.

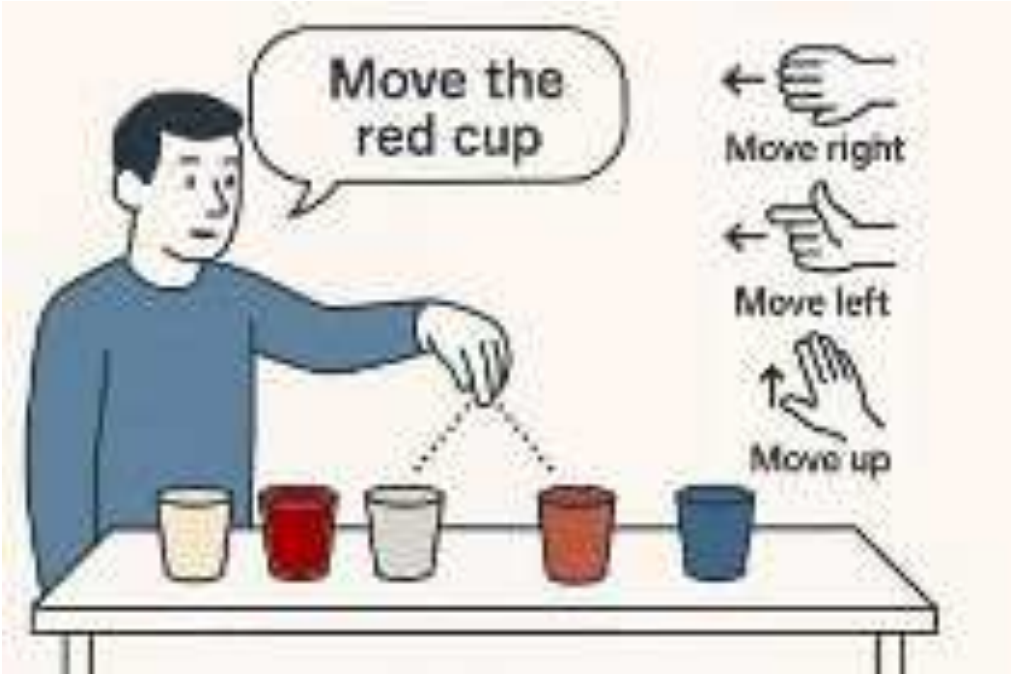
2. Interpretovat instrukce a různá omezení



2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions

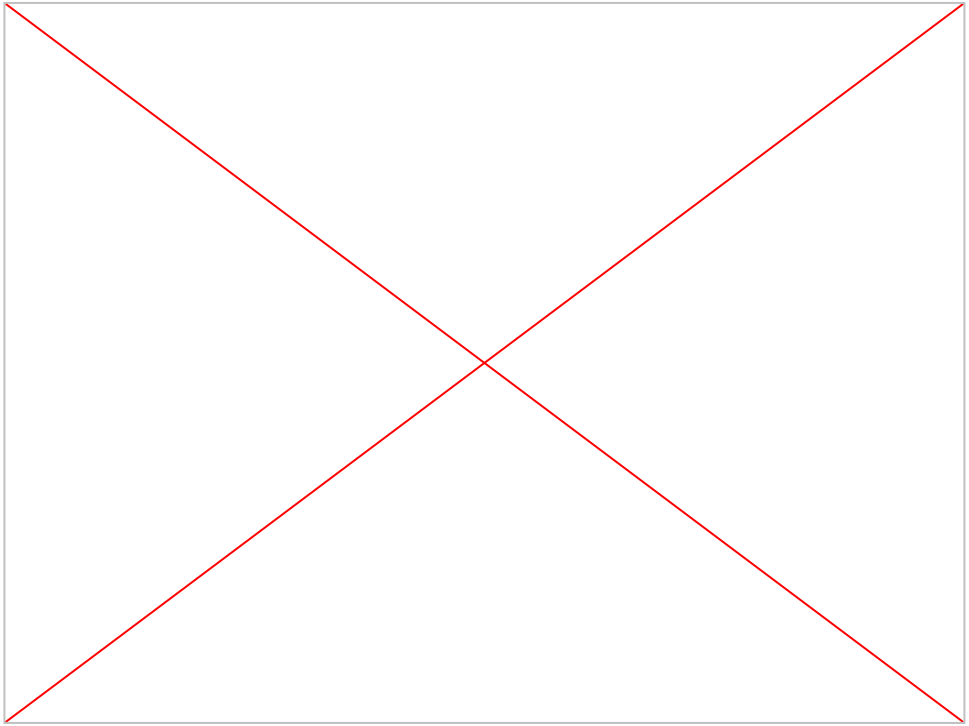
“Move this red cup”



2. Interpretovat instrukce a různá omezení

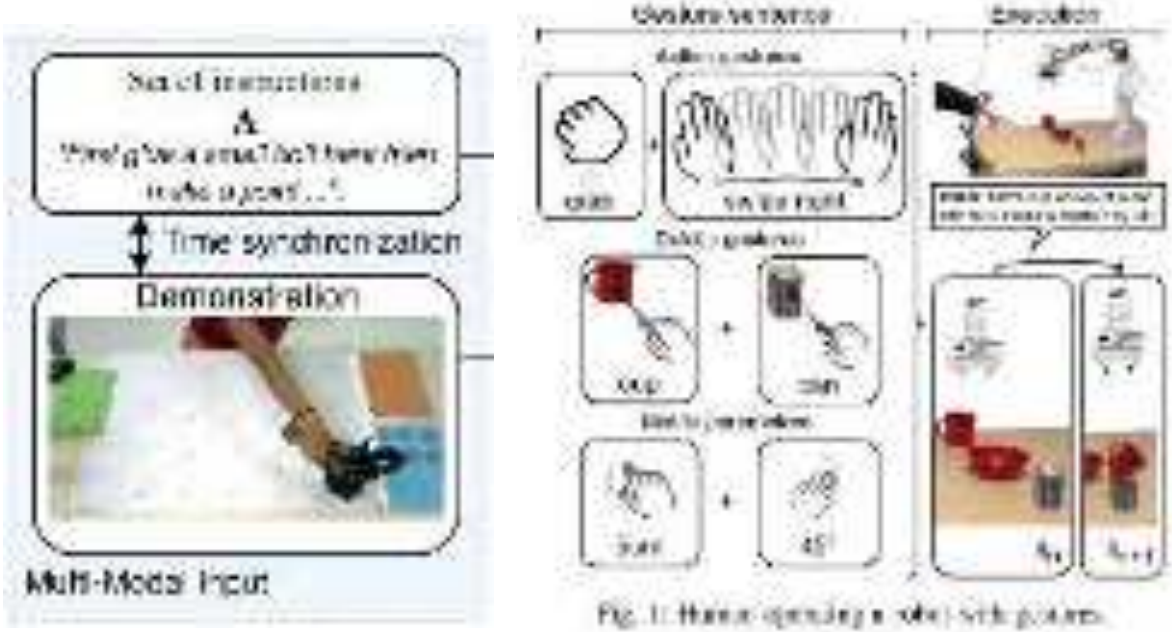
- Multimodal - ambiguity in objects and actions *“Move this red cup”*





2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*



P.Vanc, J.K.Behrens, K.Stepanova, V.Hlavac "Communicating human intent to a robotic companion by multi-type gesture sentences." *IROS*, 2023. http://imitrob.ciirc.cvut.cz/publications/chi23

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*



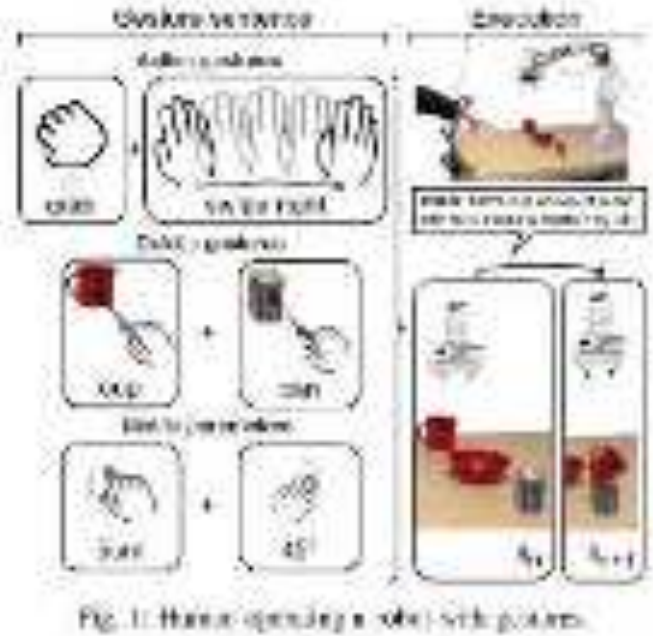


P.Vanc, J.K.Behrens, K.Stepanova, V.Hlavac "Communicating human intent to a robotic companion by multi-type gesture sentences." *IROS*, 2023. http://imitrob.ciirc.cvut.cz/publications/chi23

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*



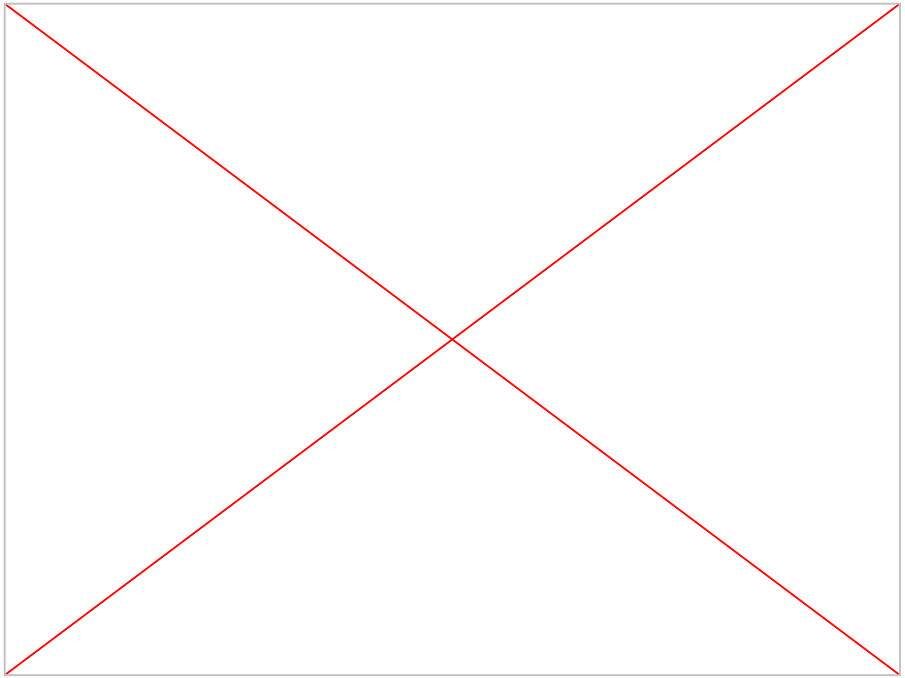




Vanc, Petr, and Karla Stepanova. "TransforMerger: Transformer-based Voice-Gesture Fusion for Robust Human-Robot Communication." arXiv preprint arXiv:2504.01708 (2025). ArXiv (2025), website, paper on ArXiv. GitHub: GitHub.

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*





Vanc, Petr, and Karla Stepanova. "TransforMerger: Transformer-based Voice-Gesture Fusion for Robust Human-Robot Communication." arXiv preprint arXiv:2504.01708 (2025). ArXiv (2025), website, paper on ArXiv. GitHub: GitHub.

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - human, task, environment - object properties, actions (SG)



2. Interpret the instructions and the rules

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - environment - object properties, actions -> SG

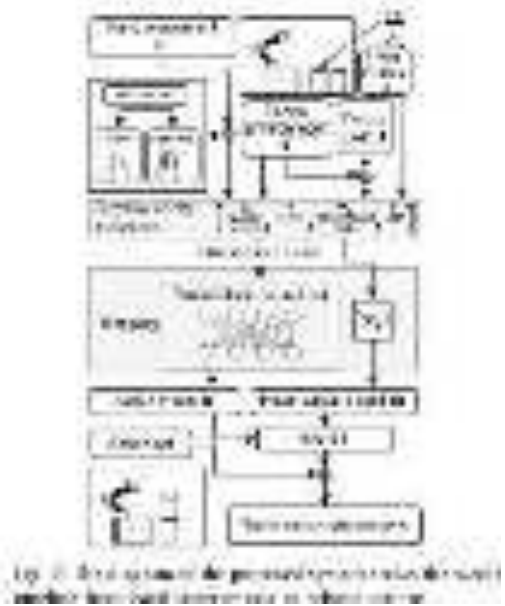




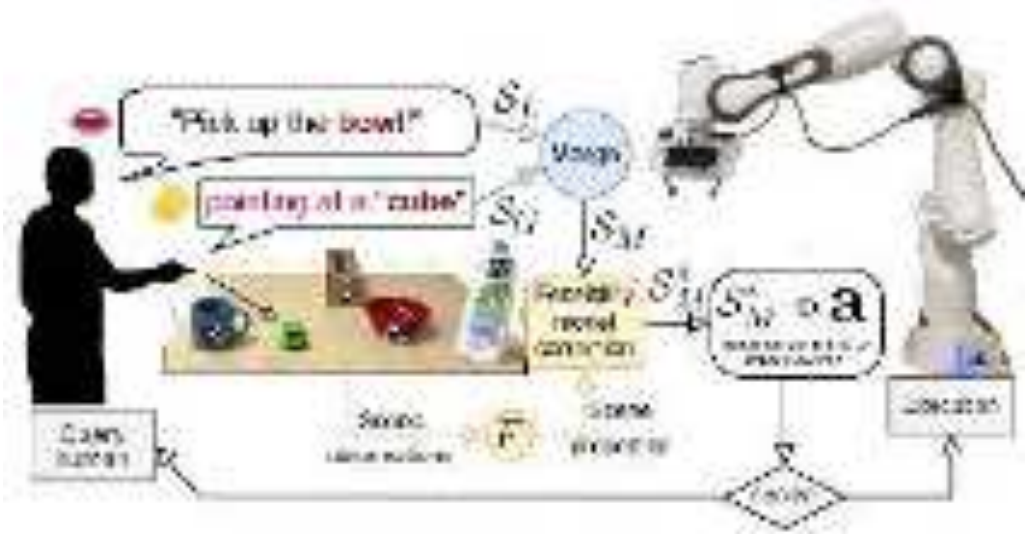
Jiang, Hanxiao, et al. "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation." *arXiv preprint arXiv:2402.15487* (2024).

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - environment, human, task



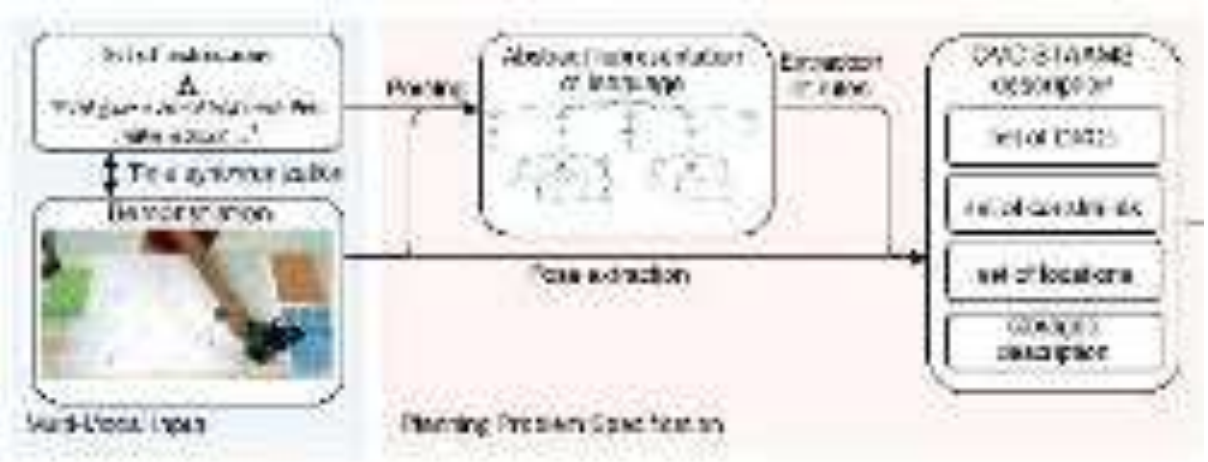
Vanc, Petr, Jan Kristof Behrens, and Karla Stepanova. "Context-aware robot control using gesture episodes." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.



Vanc, Petr, Radoslav Skoviera, and Karla Stepanova. "Tell and show: Combining multiple modalities to communicate manipulation tasks to a robot." *arXiv preprint arXiv:2404.01702* (2024).

2. Interpretovat instrukce a různá omezení

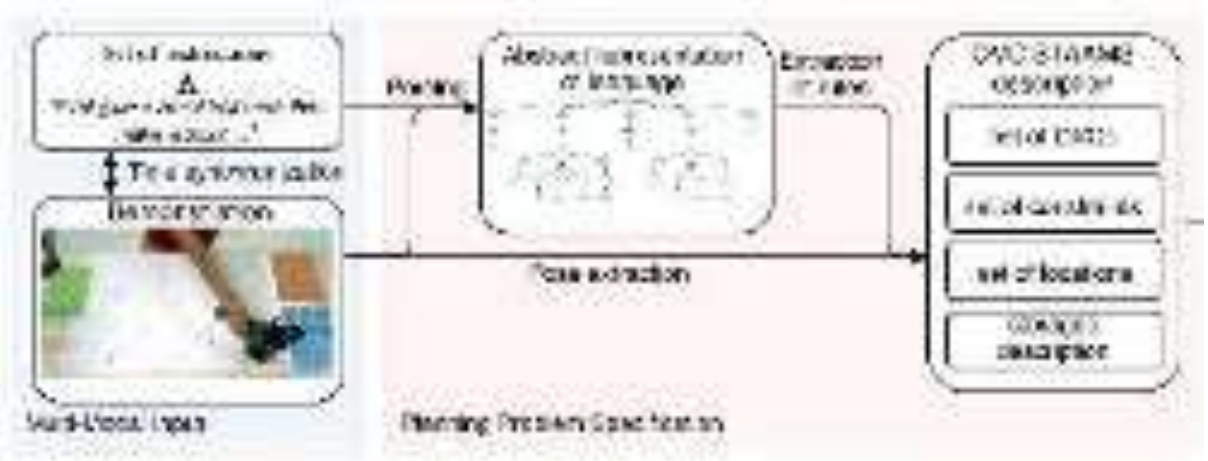
- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - scene, human, task
- Rules and preconditions - ordering and spatial



Behrens, JK, Stepanova, K., et al. "Specifying dual-arm robot planning problems through natural language and demonstration." IEEE Robotics and Automation Letters 4.3 (2019): 2622-2629. http://imitrob.ciirc.cvut.cz/planning.html

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - scene, human, task
- Rules and preconditions - ordering and spatial





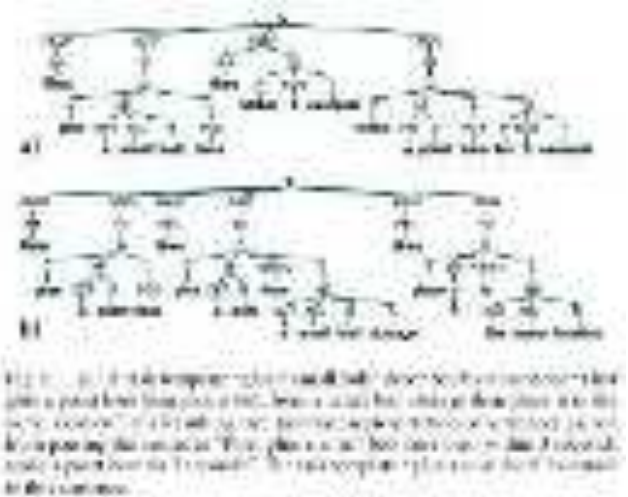
Behrens, JK, Stepanova, K., et al. "Specifying dual-arm robot planning problems through natural language and demonstration." IEEE Robotics and Automation Letters 4.3 (2019): 2622-2629. http://imitrob.ciirc.cvut.cz/planning.html

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - scene, human, task
- Rules and preconditions - ordering and spatial





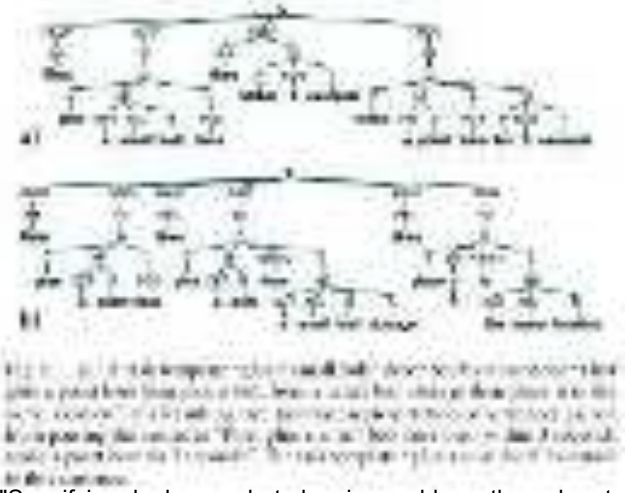


Behrens, JK, Stepanova, K., et al. "Specifying dual-arm robot planning problems through natural language and demonstration." IEEE Robotics and Automation Letters 4.3 (2019): 2622-2629. http://imitrob.ciirc.cvut.cz/planning.html

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - scene, human, task
- Rules and preconditions - ordering and spatial







Behrens, JK, Stepanova, K., et al. "Specifying dual-arm robot planning problems through natural language and demonstration." IEEE Robotics and Automation Letters 4.3 (2019): 2622-2629. http://imitrob.ciirc.cvut.cz/planning.html

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - scene, human, task
- Rules and preconditions - ordering and spatial







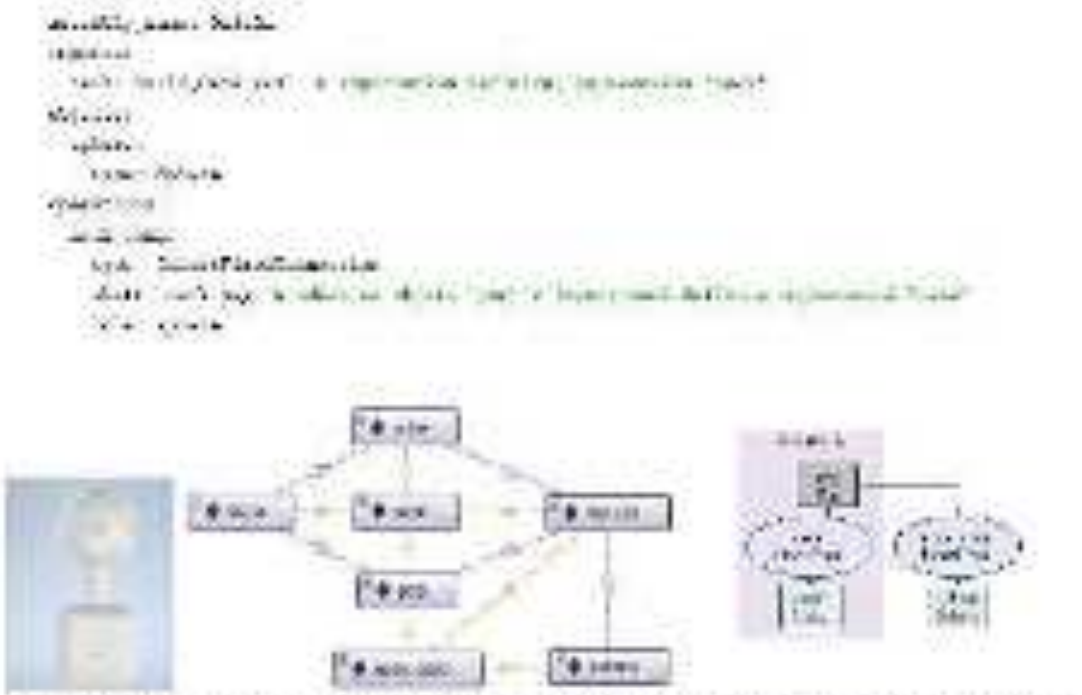
Behrens, JK, Stepanova, K., et al. "Specifying dual-arm robot planning problems through natural language and demonstration." IEEE Robotics and Automation Letters 4.3 (2019): 2622-2629. http://imitrob.ciirc.cvut.cz/planning.html

2. Interpretovat instrukce a různá omezení

- Multimodal - ambiguity in objects and actions *“Move this red cup”*
- Context - scene, human, task
- Rules and preconditions - implicit rules

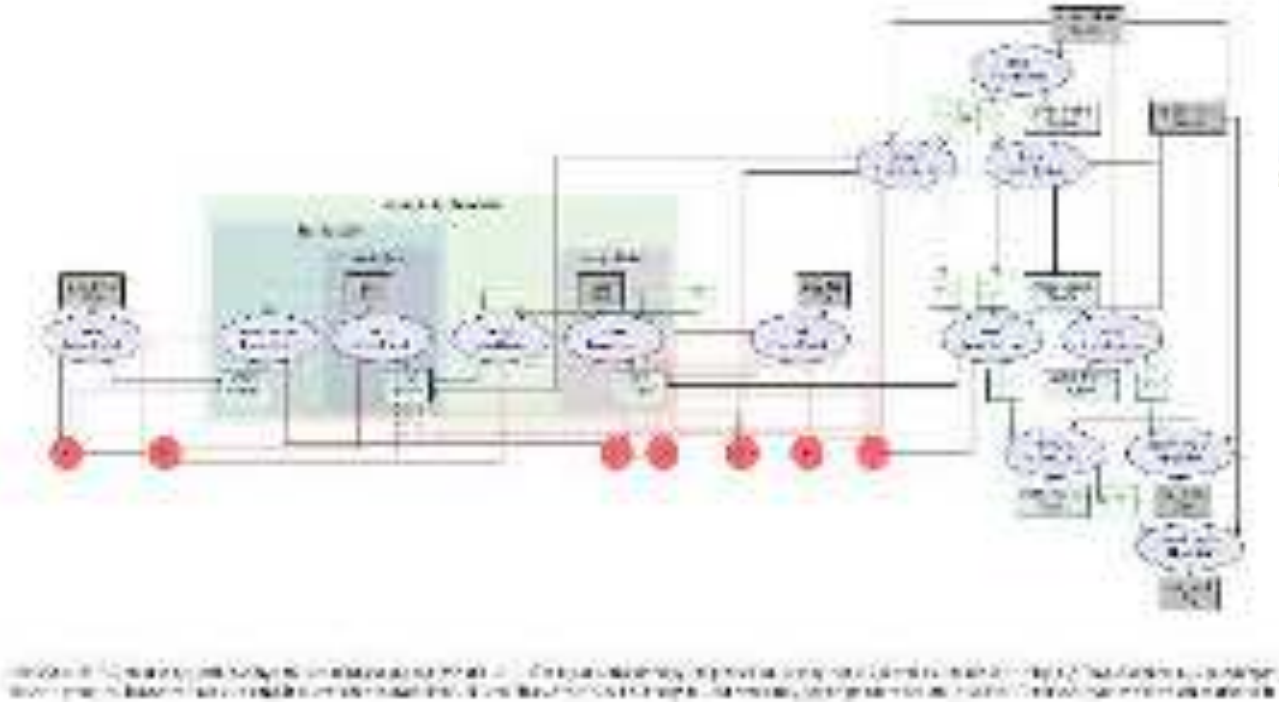


3. Reprezentovat získanou znalost



Skoviera, R., Stepanova, K., Mejdrechova M.. "Collaborative robotic workspace - task representation" (2020).

3. Reprezentovat získanou znalost







Skoviera, R., Stepanova, K., Mejdrechova M.. "Collaborative robotic workspace - task representation" (2020).

3. Reprezentovat získanou znalost



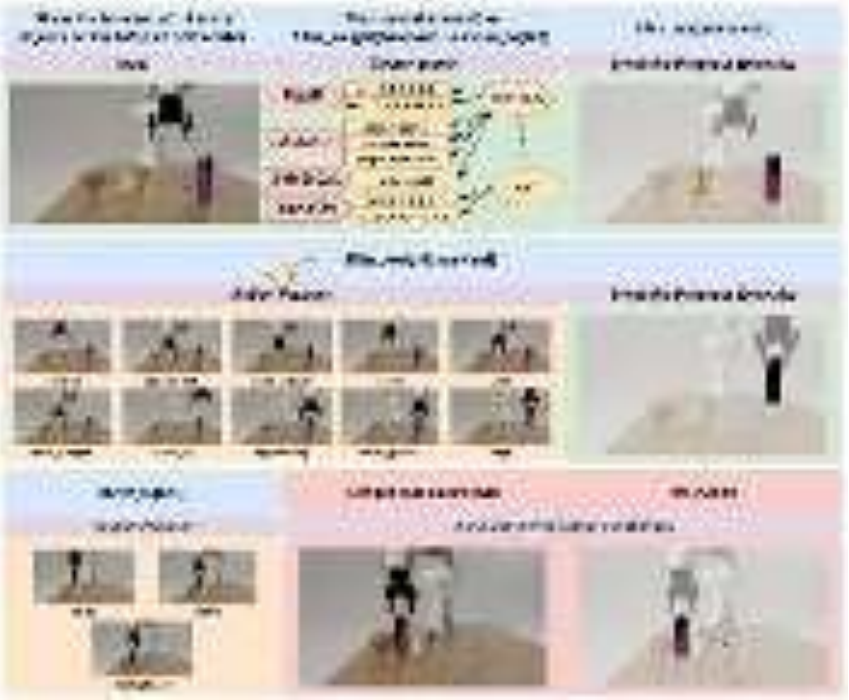
- Probabilistic knowledge graph
  - Efficient (hierarchical decomposition)
  - Easy to incorporate (link) new information
  - Variety of information
    - General “facts”
    - Task specific information
    - User preferences
  - Architecture is “classical” – deterministic “facts”, rules, validation, readable
  - Uses “AI” for inference – find connections & missing data, infer knowledge, fit new information
- Partial / missing information
  - Decide when to **explore (interactive perception)** or **ask the human**
  - **Simulation** (“think” about new situations and how to solve them)
  - Compose (more) complete knowledge from **multiple partial observations**

3. Reprezentovat získanou znalost - Spolupráce

- Planning and Scheduling
  - User actions – alternative plans, fast replanning
  - User preferences – e.g. what to do when
  - Long term reasoning – key aspect of autonomy
- Safety
  - Addition to physical safety
  - Recognition of anomalous (potentially problematic / dangerous) situations

4. Hledat chybějící informace/upřesňovat si





Nazarczuk, M., Behrens, J. K., Stepanova, K., Hoffmann, M., & Mikolajczyk, K. (2024). Closed Loop Interactive Embodied Reasoning for Robot Manipulation." *ICRA* (2025), website, paper on ArXiv. GitHub.

Michal Nazarczuk[1]
Jan Kristof Behrens[2]
Karla Stepanova[2]
Matej Hoffmann[2]
Krystian Mikolajczyk[1]
1) Imperial College London
2) Czech Technical University



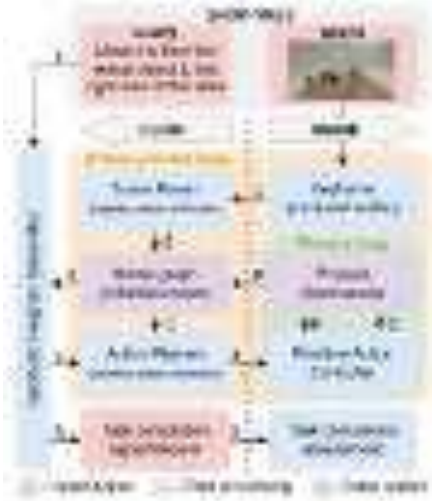
CLIER

- Real world execution
- Common failures
- Recovery capabilities

Closed Loop Interactive Embodied Reasoning

4. Hledat chybějící informace/upřesňovat si

Closed-loop exploration of the scene to refine the knowledge in the scene graph



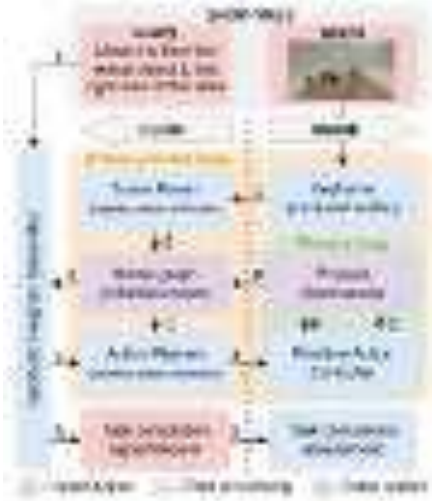


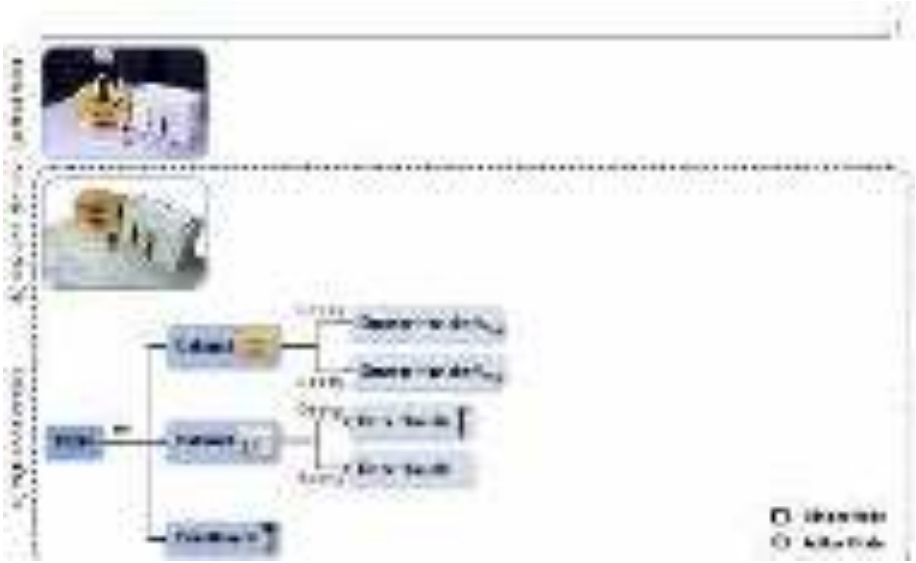
Nazarczuk, M., Behrens, J. K., Stepanova, K., Hoffmann, M., & Mikolajczyk, K. (2024). Closed Loop Interactive Embodied Reasoning for Robot Manipulation." *ICRA* (2025), website, paper on ArXiv. GitHub.

Jiang, Hanxiao, et al. "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation." *arXiv preprint arXiv:2402.15487* (2024).

4. Hledat chybějící informace/upřesňovat si

Closed-loop exploration of the scene to refine the knowledge in the scene graph



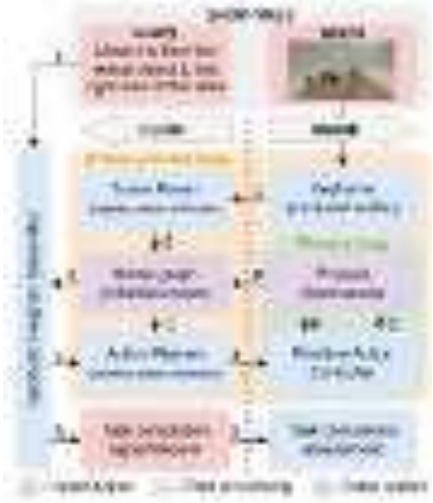


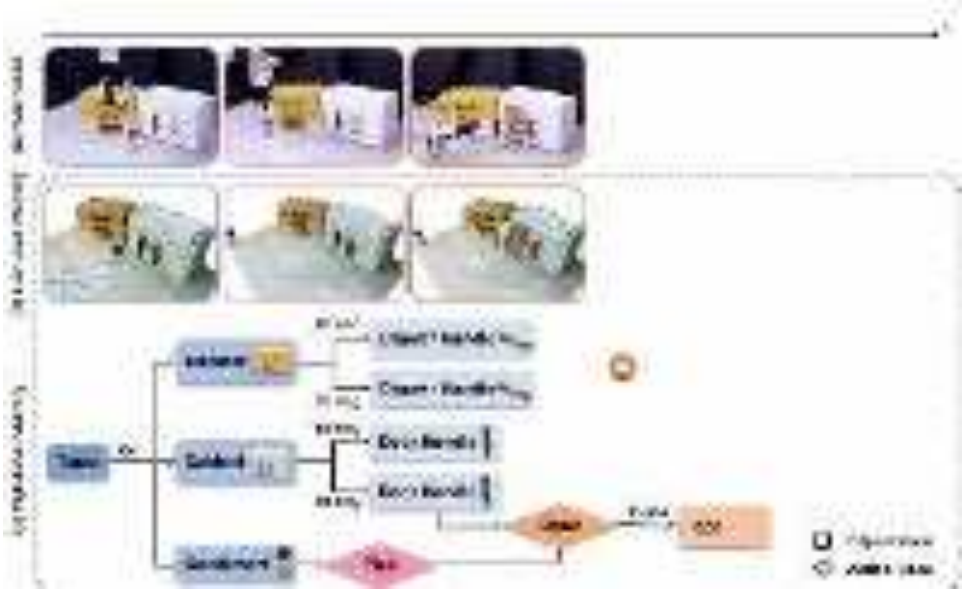
Nazarczuk, M., Behrens, J. K., Stepanova, K., Hoffmann, M., & Mikolajczyk, K. (2024). Closed Loop Interactive Embodied Reasoning for Robot Manipulation." *ICRA* (2025), website, paper on ArXiv. GitHub.

Jiang, Hanxiao, et al. "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation." *arXiv preprint arXiv:2402.15487* (2024).

4. Hledat chybějící informace/upřesňovat si

Closed-loop exploration of the scene to refine the knowledge in the scene graph



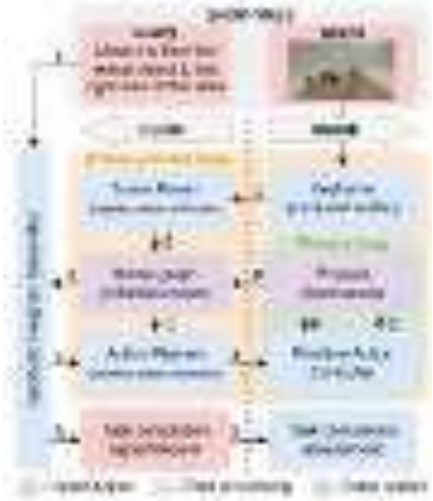


Nazarczuk, M., Behrens, J. K., Stepanova, K., Hoffmann, M., & Mikolajczyk, K. (2024). Closed Loop Interactive Embodied Reasoning for Robot Manipulation." *ICRA* (2025), website, paper on ArXiv. GitHub.

Jiang, Hanxiao, et al. "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation." *arXiv preprint arXiv:2402.15487* (2024).

4. Hledat chybějící informace/upřesňovat si

Closed-loop exploration of the scene to refine the knowledge in the scene graph



Nazarczuk, M., Behrens, J. K., Stepanova, K., Hoffmann, M., & Mikolajczyk, K. (2024). Closed Loop Interactive Embodied Reasoning for Robot Manipulation." *ICRA* (2025), website, paper on ArXiv. GitHub.



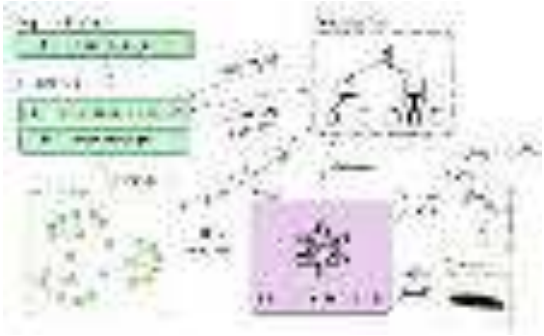
Jiang, Hanxiao, et al. "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation." *arXiv preprint arXiv:2402.15487* (2024).

4. Hledat chybějící informace/upřesňovat si









Nazarczuk, M., Behrens, J. K., Stepanova, K., Hoffmann, M., & Mikolajczyk, K. (2024). Closed Loop Interactive Embodied Reasoning for Robot Manipulation." *ICRA* (2025), website, paper on ArXiv. GitHub.

Mark Horpynych, Interactive Robotic Scene Exploration Using Vision-Language Models, Bachelor thesis, 2025.

5. Vykonat úlohu - reaktivně a bezpečně

- Reactive and closed-loop

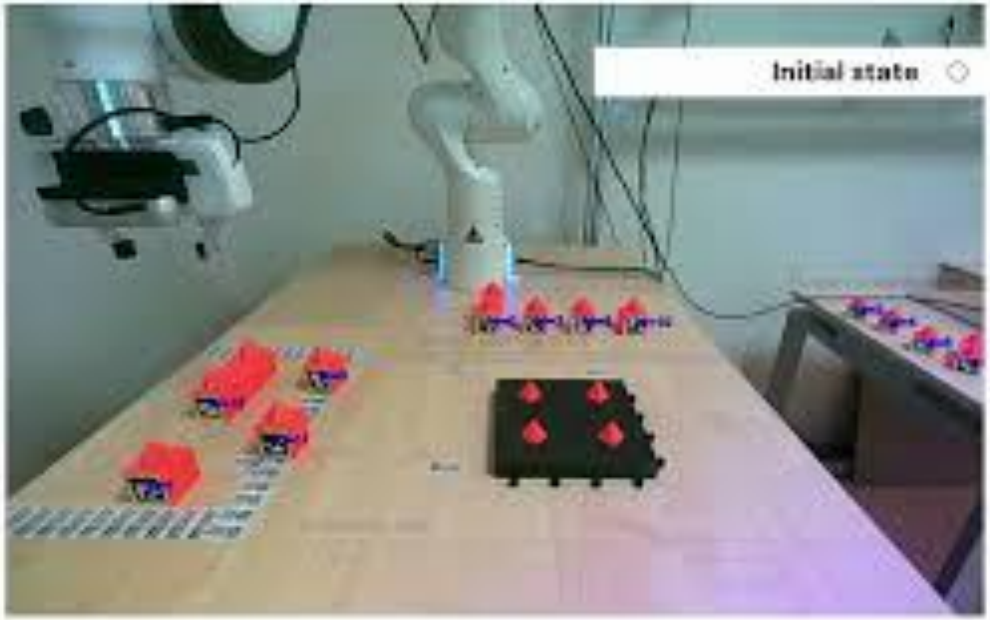






5. Vykonat úlohu - reaktivně a bezpečně

- Reactive and closed-loop

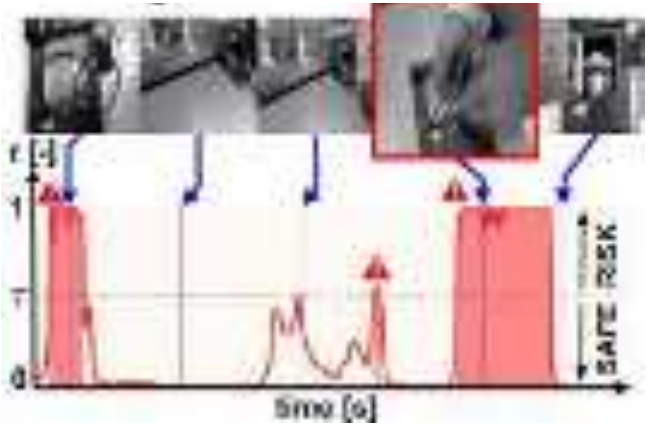


We evaluated our solution using a setup involving a Franka Emika Panda robot and a human tracked using HTC Vive VR gloves, where task assignment options depend on the manipulated object's location.

5. Vykonat úlohu - reaktivně a bezpečně

- Reactive and closed-loop for changes on any level
- Safe execution - use senses



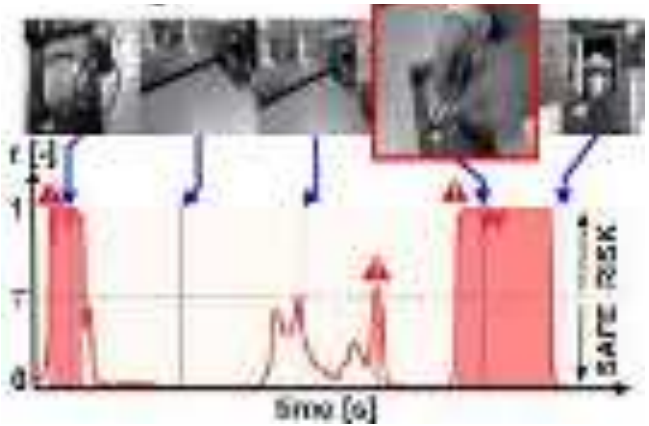


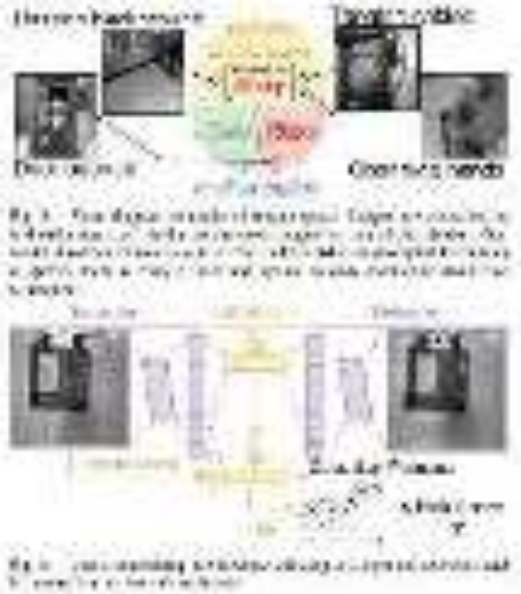
Vanc, P., Franzese, G., Behrens, J. K., Della Santina, C., Stepanova, K., & Kober, J.. "ILeSiA: Interactive Learning of Situational Awareness from Camera Input". arXiv preprint arXiv:2409.20173 (2024)., website, paper on ArXiv. GitHub: GitHub.

5. Vykonat úlohu - reaktivně a bezpečně

- Reactive and closed-loop for changes on any level
- Safe execution





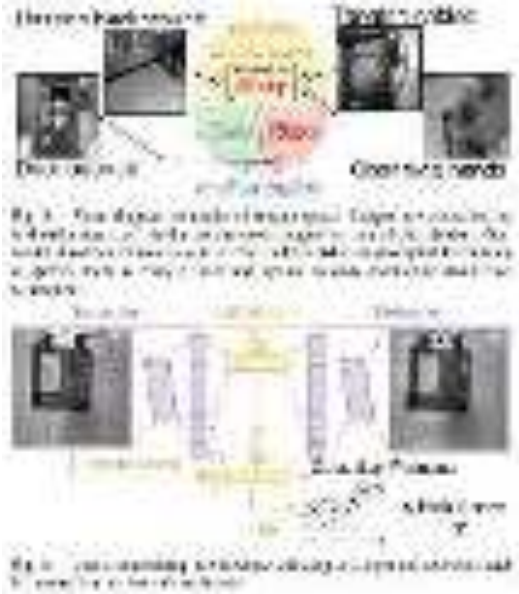


Vanc, P., Franzese, G., Behrens, J. K., Della Santina, C., Stepanova, K., & Kober, J.. "ILeSiA: Interactive Learning of Situational Awareness from Camera Input". arXiv preprint arXiv:2409.20173 (2024)., website, paper on ArXiv. GitHub: GitHub.

5. Vykonat úlohu - reaktivně a bezpečně

- Reactive and closed-loop for changes on any level

ILeSiA: Interactive Learning of Situational Awareness from Camera Input



Vanc, P., Franzese, G., Behrens, J. K., Della Santina, C., Stepanova, K., & Kober, J.. "ILeSiA: Interactive Learning of Situational Awareness from Camera Input". arXiv preprint arXiv:2409.20173 (2024)., website, paper on ArXiv. GitHub: GitHub.

5. Vykonat úlohu - reaktivně a bezpečně

- Reaktivní vykonávání na různých úrovních
- Bezpečné vykonání úlohy
- Zobecnitelné na nové roboty, prostředí, úlohy - znovupoužitelné podúlohy, hledání pravidel
  - AI doesn’t keep the memory well
  - How to use prior knowledge in the new environment?

Proč je spolupráce a komunikace s člověkem potřeba?

- Sdílený prostor
- Člověk vykonává část úloh, robot jinou
- Člověk nese znalosti o úloze - potřeba komunikace
  - Implicitní expertní znalost - svět není celý popsaný
  - Člověk nese záměr úlohy
- Spolupráce a komunikace umožňuje adaptabilitu robota
- Spolupráce velmi zrychlí učení
- Člověk důvěřuje systému, který je vysvětlitelný a reaktivní
  - Spolupráce vs. bezpečnost
    - Dříve byla bezpečnost založená na oddělení člověka a robota.
    - Dnes je založená na porozumění, spolupráci a adaptaci.

Závěrem

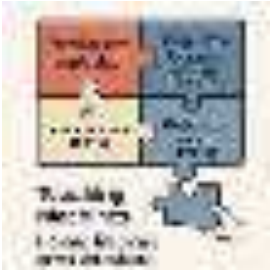
- Aby roboti dokázali skutečně spolupracovat s lidmi, musí rozumět významu toho, co dělají – nestačí jen rozpoznat slova nebo objekty
- Klíčové je kombinovat více způsobů vnímání: zrak, interakci, řeč…
- To, jak si robot „ukládá“ úlohu a znalosti snadno selhává při změně situace
- K flexibilním a spolupracujícím robotům máme stále ještě kus cesty
-

Závěrem

- Aby roboti dokázali skutečně spolupracovat s lidmi, musí rozumět významu toho, co dělají – nestačí jen rozpoznat slova nebo objekty
- Klíčové je kombinovat více způsobů vnímání: zrak, interakci, řeč…
- To, jak si robot „ukládá“ úlohu a znalosti snadno selhává při změně situace
- K flexibilním a spolupracujícím robotům máme stále ještě kus cesty
- Nejdřív je třeba přidat do AI více modalit a pak…
-

Závěrem

- Aby roboti dokázali skutečně spolupracovat s lidmi, musí rozumět významu toho, co dělají – nestačí jen rozpoznat slova nebo objekty
- Klíčové je kombinovat více způsobů vnímání: zrak, interakci, řeč…
- To, jak si robot „ukládá“ úlohu a znalosti snadno selhává při změně situace
- K flexibilním a spolupracujícím robotům máme stále ještě kus cesty
- Nejdřív je třeba přidat do AI více modalit a pak…
- …umožnit kombinovat více způsobů učení a zahrnutí předchozí informace
- …a poučit se z dalších věd
  - Psychologie - jak evaluovat a navrhnout *experimenty a učení mezi člověkem a robotem*
  - Kognitivní věda pro modely *jak lidé vytvářejí, strukturují a zobecňují koncepty*
  - AI pro nástroje, abychom *tyto kognitivní koncepty mohli implementovat (učení se vnitřní reprezentace)*
  - Neurovědu pro *vhledy jak je vnitřní reprezentace stavěná a upravovaná v biologickém systému*
  - Robotiku, aby AI reprezentace skutečně mohly něco dělat a vnímat v reálném světě.



ELLIOT: European Large Open Multi-Modal Foundation Models For Robust Generalization On Arbitrary Data Streams

ELLIOT je 4-letý EU projekt, který se snaží vytvořit důvěryhodný **evropský multimodální obecný model**, který bude nasaditelný v průmyslu.

- Robotická práce s povrchem (natírání, sváření, postřik, broušení…)
  - Robot se učí úlohu přímo od člověka, ne složitým programováním
  - Člověk robota vede pomocí domonstrace, slov nebo gest
  - Robot vnímá tvar a polohu objektů v prostoru (3D vnímání)
  - Pohyby robota se automaticky přizpůsobují skutečnému tvaru součástí
  - Pokud se objekt změní nebo posune, robot se dokáže přizpůsobit
  - Robot průběžně kontroluje kvalitu své práce pomocí kamer
  - Člověk může robota kdykoli opravit nebo zastavit
  - Robot funguje jako pomocník, ne jako náhrada člověka

Spolupracující firma: Robotwin, s.r.o.





https://elliot-ai.eu/project-overview

Personalizace a budoucnost



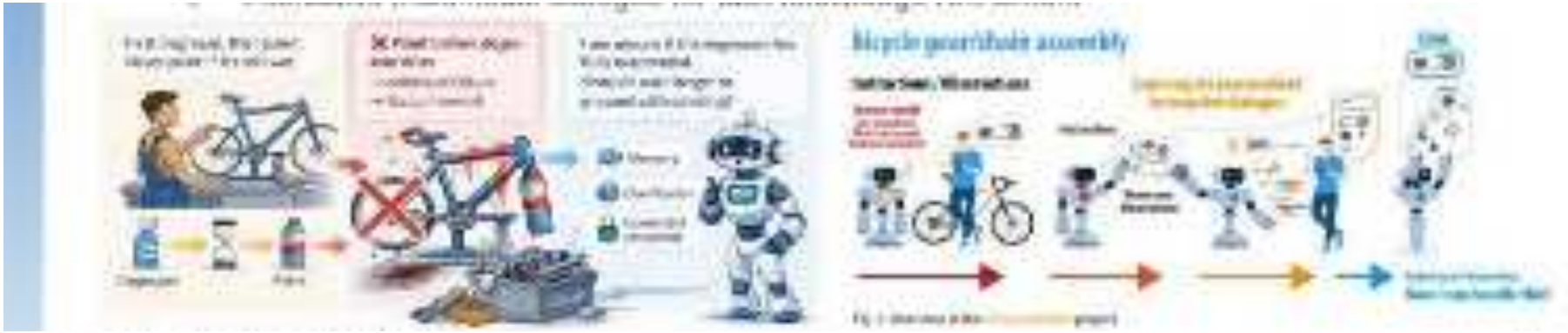




PersonalRobot: Personalizace robotů pomocí multimodálního interaktivního dialogu člověka s robotem

Cílem projektu je, aby si robot byl schopný zapamatovat minulé instrukce, rozhodovat se v situacích s neurčitostí a vědět, kdy neví dost a měl by se tedy doptat na informace člověka, přičemž bere do úvahy jeho aktuální stav (vytížení, pozornost…) i kontext prostředí

- Úprava znalostí podle nových pozorování
- Kontext úlohy a člověka
- Nalezení chybějících znalostí v paměti robota
- Interaktivní dialog pro vylepšení znalostí



PersonalRobot: Customizing Robots via Multimodal Interactive Human-Robot Dialogues

The goal is a robot that remembers past interactions, reasons about uncertainty, and knows when it doesn't know enough, while taking into account the context of the human (attention, mental load, preferences) and environment (available tools, etc.)

- Four main subtasks
  - Updating task knowledge from multimodal partial observations
  - Context - modeling human and environment
  - Identifying knowledge gaps in a robot memory system
  - Interactive multimodal dialogue for task knowledge refinement



