AI Robotics Integration

- **A Generalist Agent [`arXiv 2022`]** *Scott Reed, et al.* (arXiv) (pdf) (Citation: 381)
- **ChatGPT for Robotics Design Principles and Model Abilities [`arXiv 2023`]** *Sai Vemprala, Rogerio Bonatti, Arthur Bucker, Ashish Kapoor* (arXiv) (pdf) (Citation: 62)
- **RT-1 Robotics Transformer for Real-World Control at Scale [`arXiv 2022`]** *Anthony Brohan, et al.* (arXiv) (pdf) (Citation: 107)
- **RT-2 Vision-Language-Action Models Transfer Web Knowledge to Robotic Control [`arXiv 2023`]** *Anthony Brohan, et al.* (arXiv) (pdf) (Citation: 6)
- **VIMA General Robot Manipulation with Multimodal Prompts [`ICML 2023`]** *Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, Linxi Fan* (arXiv) (pdf) (Citation: 53)
- **PaLM-E An Embodied Multimodal Language Model [`arXiv 2023`]** *Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, Pete Florence* (arXiv) (pdf) (Citation: 194)

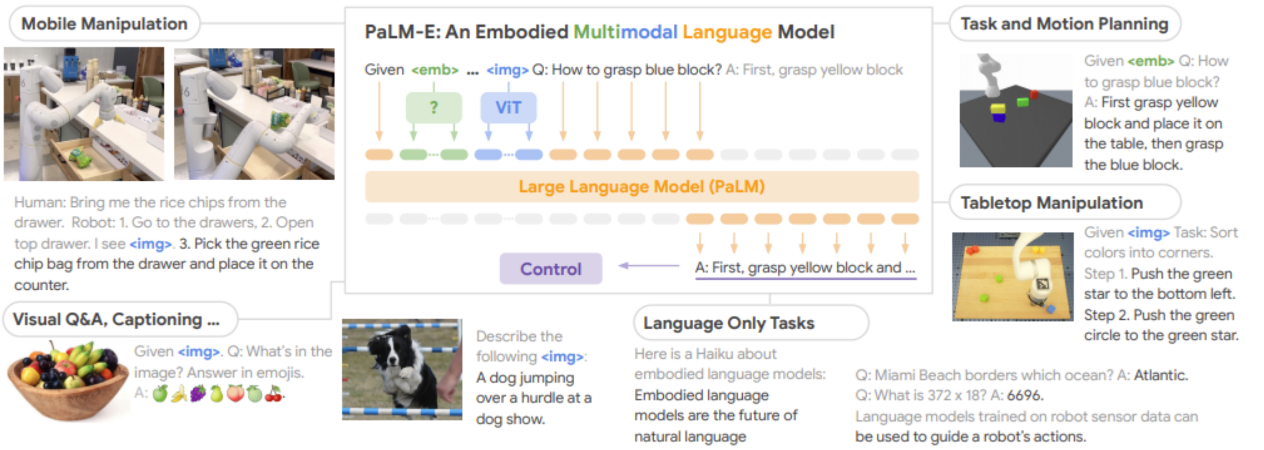
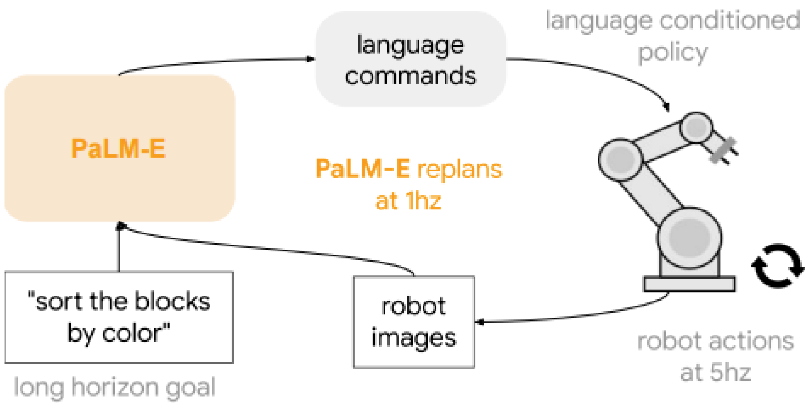


Figure 1: PaLM-E is a single general-purpose multimodal language model for embodied reasoning tasks, visual-language tasks, and language tasks. PaLM-E *transfers* knowledge from visual-language domains into embodied reasoning – from robot planning in environments with complex dynamics and physical constraints, to answering questions about the observable world. PaLM-E operates on *multimodal sentences*, i.e. sequences of tokens where inputs from arbitrary modalities (e.g. images, neural 3D representations, or states, in green and blue) are inserted alongside text tokens (in orange) as input to an LLM, trained end-to-end.

  - An Embodied Model with Multimodal Inputs
    - Multimodal inputs include visual, **continuous state estimation**, and textual input encodings.
    - Train the model end-to-end, with a pre-trained LLM
    - Tasks: sequential robotic manipluation planning, visual question answering, and captioning
    - Establish the link between words and percepts
  - Dataflow
    - Inputs (images and state estimates) -> tokens (image, state estimates, and language has the same dimension) -> self-attention layers (decoder-only LLM) -> Output (**sequential**

decisions in natural text) -> Low-level policy or planner (translate language decisions into actions)



- Terminologies
  - $w_i\in \mathcal{W}$ -> a token
  - I -> an image
  - $\mathcal{X} \in \mathbb{R}^k$ -> embedding space (if a token's length is 3, then the embedding space is a 3-dimensional space)
  - $\gamma : \mathcal{W} \rightarrow \mathcal{X}$ -> **Text encoder**: a LLM that embeds a token w_i into a word token embedding space. PaLM model is employed in this work.
  - $\phi:\mathcal{O} \rightarrow \mathcal{X}^q$ -> **Vision encoder**: an encoder ϕ that maps a continuous observation space $\mathcal{O}$ into a sequence of q-many vectors in $\mathcal{X}$. ViT-22B and ViT-4B are employed in this work.
  - ϕ_{state} -> State estimation encoder (a MLP that maps inputs into the language embedding space).
  - $D = \left{\left(I_{1: u_i}^i, w_{1: L_i}^i, n_i\right)\right}*{i=1}^N$ -> Training dataset that contains u_i images, a text $w*{1:L_i}$, and an index n_i.
    - u_i images -> Continuous observations
    - $w_{1: L_i}^i$ -> Text contains a prefix part (i.e., prompt template) formed from multi-modal sentences and a prediction target that only contains text tokens.