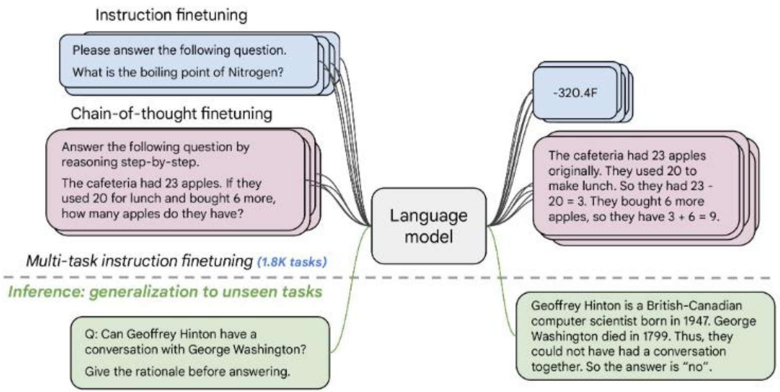
VIMA: General Robot Manipulation with Multimodal Prompts.

Presenter: Abhiram Maddukuri

10/19/2023

Motivation: Multi-task Success for LM's

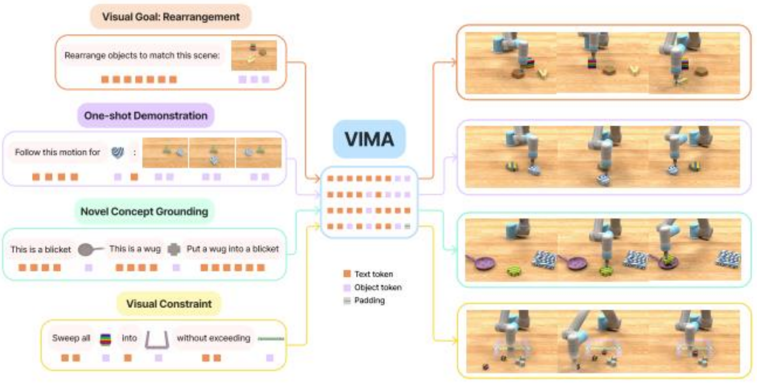
- Language Models very successful in generalizing to multi-task settings
- Specifically effective in zero-shot situations when prompted a problem
- How can we extend this to robotics agents?



Raffel et al., 2020

Motivation/Problem: Multimodal Task Specification

- Many ways to specify tasks for robots
  - Natural language
  - Imitation video
  - Text interleaved with language



- Previous works employ different architectures, objectives, data pipelines, etc. for different tasks or only take as unimodal input

- Key challenge is to encode and consume these prompts in a unified way

Context / Related Work / Limitations of Prior Work

- Multi-task/Multimodal/zero-shot learning
  - Raffel et al., 2020; Alayrac et al.,2022; Reed et al., 2022
  - Does not unify all three for robotic tasks
- Transformer-based agents
  - Chen et al., 2021; Janner et al., 2021; Brohan et al., 2022
  - Does not consider multimodal prompt-based learning

Problem

- Observation Space: O
- Action Space: A
- Input: Prompt P of length l, where $P = [x_{1, \dots,} x_n], x_i \in \{text, \ image\}$
- Goal: Learn a policy $\pi_\theta(a_t \mid P, H)$, where $H = [o_{1,} a_{1, \dots,} o_{t-1}, a_{t-1}], \ (o_i \ \in O, \ a_i \ \in A)$

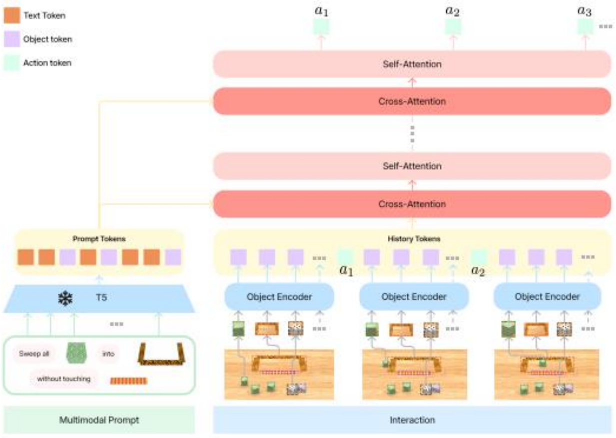
Proposed Approach: Prompt Encoding

- 4 Aspects to individually tokenize for unified encoding
    - Text input: T5 Tokenization
    - Full scene input: Extract and crop objects via Faster-RCNN, encode bounding box position with ViT, further encode with MLP
    - Specific object input: Same as full scene but with dummy bounding boxes
    - Imitation video input: Use keyframes

- After tokenization, feed into T5

Proposed Approach: Policy

- Encoder-Decoder architecture
    - Decoder alternates between cross-attention with encoded tokens and self-attention with history input. L such layers.
    - Decoder outputs action at each time step



- Behavioral cloning objective
    - Loss function: $\sum_{t=1}^{T} -\log \pi_\theta(a_t \mid P, H)$

Experimental Setup: Domain

- All experiments done in the Ravens Simulator (Zeng et al., 2020)
- 6 categories of task suites:
  - Object manipulation, visual goal reaching, novel concept grounding, one-shot video imitation, visual constraint satisfaction, visual reasoning
  - Mostly “pick and place” and “wipe” tasks

Experimental Setup: Domain continued

- Experiments evaluate varying levels of zero-shot generalization
    - L1: Placement Generalization
    - L2: Combinatorial Generalization
    - L3: Novel Object Generalization
    - L4: Novel Task Generalization

Experimental Setup: Baselines

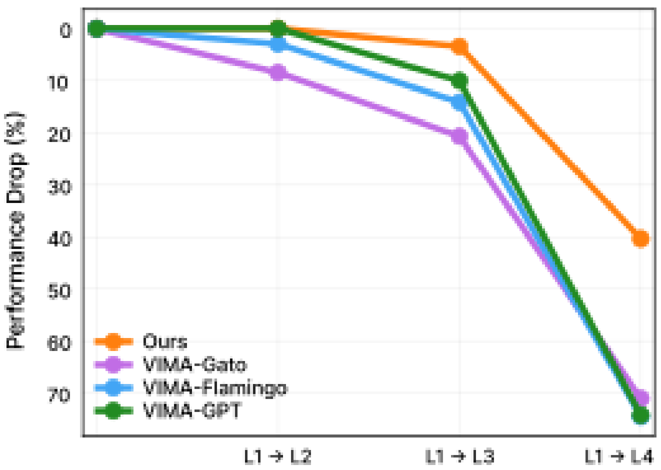
- No prior method that works with multimodal prompting; have to repurpose other models
    - VIMA-Gato
    - VIMA-Flamingo
    - VIMA-GPT
    - None of the models use cross-attention

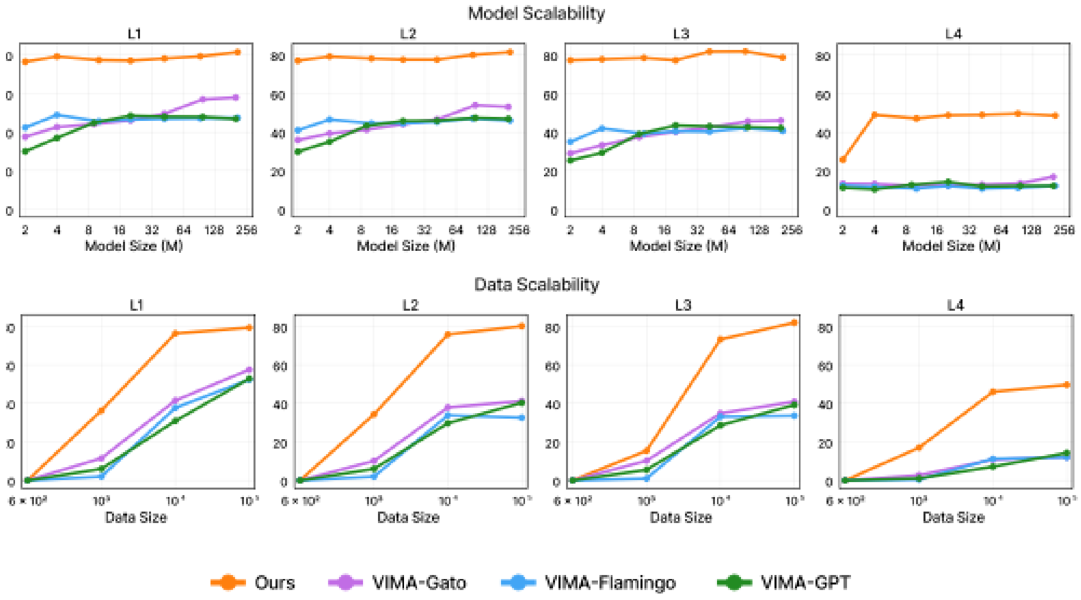
Experimental Setup

- Want to test 3 things
  - Evaluate the most important components in multi-task transformer agents
  - Scaling properties: data and model size
  - Ablations

Results: Generalization

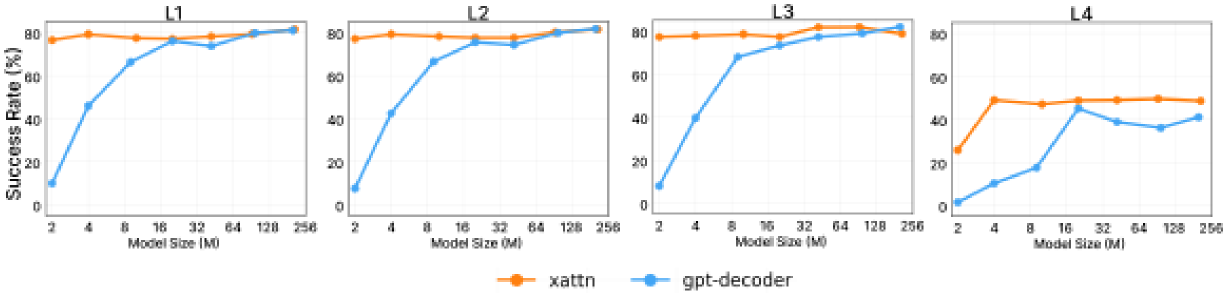
- VIMA shows best results in terms of generalization and scalability

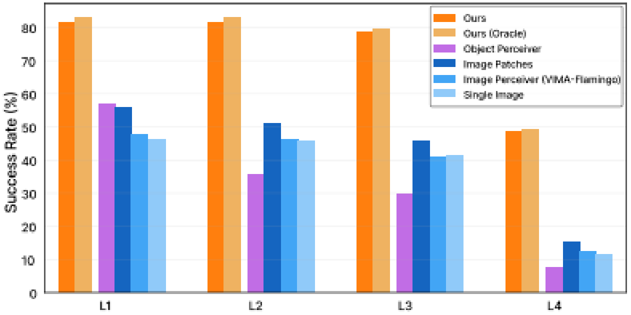




Results: Ablations

- Ablations show the importance of the cross-attention and object detection

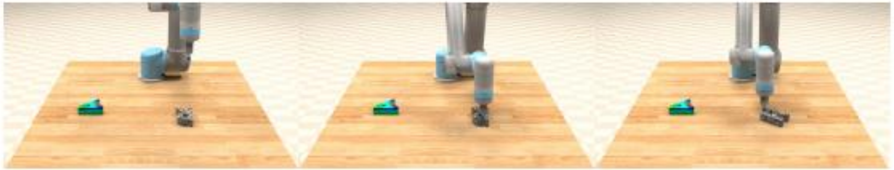




Example Trajectory: Manipulation

Task 03: Rotate objects clockwise by certain degrees along z-axis. Only rotationally asymmetric objects are considered in this task.

- **Prompt:** `Rotate the {object}`$_1$ `{angles} degrees.`
- **Description:** The agent is required to rotate all objects in the workspace specified by the image placeholder `{object}`$_1$. There are also objects with different color-shape combinations in the workspace as distractors. `{angles}` is the sampled degree that needs to be rotated. A target angle is sampled from $30°$, $60°$, $90°$, $120°$, and $150°$.
- **Success Criteria:** The position of the specified object matches its original position, and the orientation matches the orientation after rotating specific angles.
- **Oracle Trajectory:** Shown in Fig. A.5 with its multimodal prompt.



Rotate the  120 degrees.

Figure A.5: Simple Object Manipulation: Task 03

Example Trajectory: Imitation

<u>**Task 10:**</u> Follow motions for specific objects.

- **Prompt:** `Follow this motion for {object}: {frame}`$_1$`...{frame}`$_i$`...{frame}`$_n$`.`
- **Description:** Image placeholder `{object}` is the target object to be manipulated and {`{frame}`$_i$} is set of workspace-like scene placeholders to represent a video trajectory, where n is the trajectory length. There is an object spawned at the center in both the workspace and the prompt video but with different textures as a distractor. The initial position of the target object matches that in `{frame}`$_1$.
- **Success Criteria:** In each step, the pose of the target object matches the pose in the corresponding video frame. Incorrect manipulation sequences are considered as failures.
- **Oracle Trajectory:** Shown in Fig. A.12 with its multimodal prompt.

Figure A.12: One-shot video imitation: Task 10

Discussion of Results

- VIMA does have noticeably outperform baselines on a diverse set of tasks and at levels of generalization and scaling
  - Importance Cross-attention and object detection emphasized in ablations
- Caveats
  - Baselines were originally intended for different tasks and were repurposed
  - It is unsurprising that using detected objects as opposed to raw pixels will be more sample efficient

Limitations

- Reliance on object detection model
    - Gives more of an advantage to VIMA, also limits its in-the-wild usability

- Simple tasks/no evaluation of long horizon tasks
- Object-centric tokens in prompt
    - Could limit general usability; ex: requirement of keyframes for one-shot imitation from video

Future Work/Directions

- More extensive evaluation in terms of domains and tasks
  - More simulators, Real robot evaluation, long horizon tasks
- Evaluation VIMA-Bench with representation learning models
  - Leverage VIMA-Bench’s comprehensiveness in testing generalization

Extended Readings

- PaLM-E (Driess et al., 2023) - repurposes the PaLM language model for robotic manipulation via multimodal prompting

- Foundation Models for Decision Making: Problems, Methods, and Opportunities (Yang et al., 2023) - survey on the state of foundation models in decision making problems

- RT-2 (Brohan et al., 2023): trains a multimodal robotic foundation model on large-scale internet data and robot trajectory data

Summary

- No unified framework for a multimodal task specification for robots
- Task specification for generalist robots are best given in a multimodal fashion; however, it is nontrivial to represent and learn across multiple modalities
- Previous work either is unimodal input/single-task or is not specified to robotics
- Authors show that cross-attention and object level-tokenization are crucial for given task
- Authors also offer VIMA-Bench, a comprehensive simulation dataset testing multitask learning and zero-shot generalization

Thank you!