VIMA Learning Poster

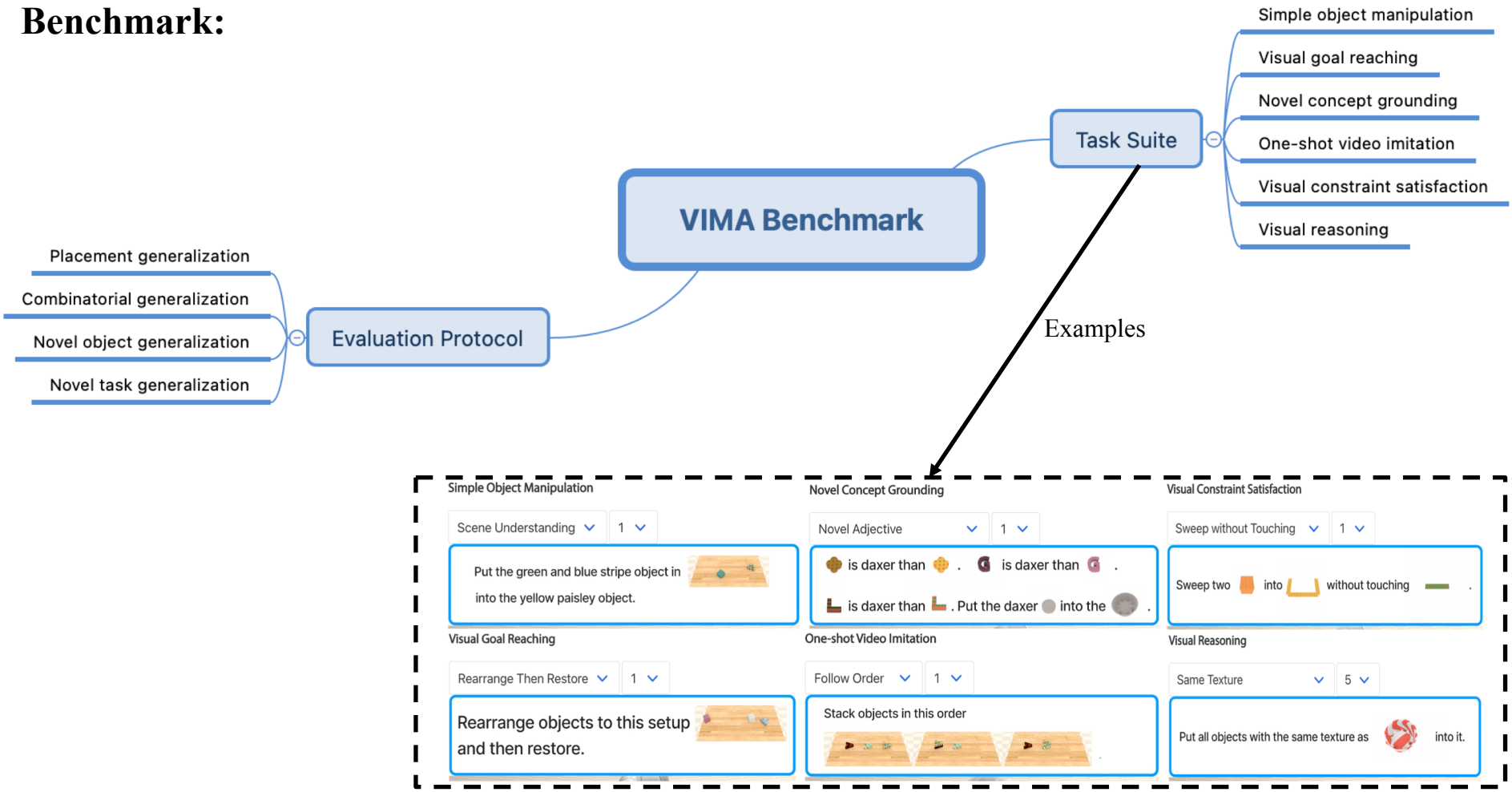
Tags: Robot Learning

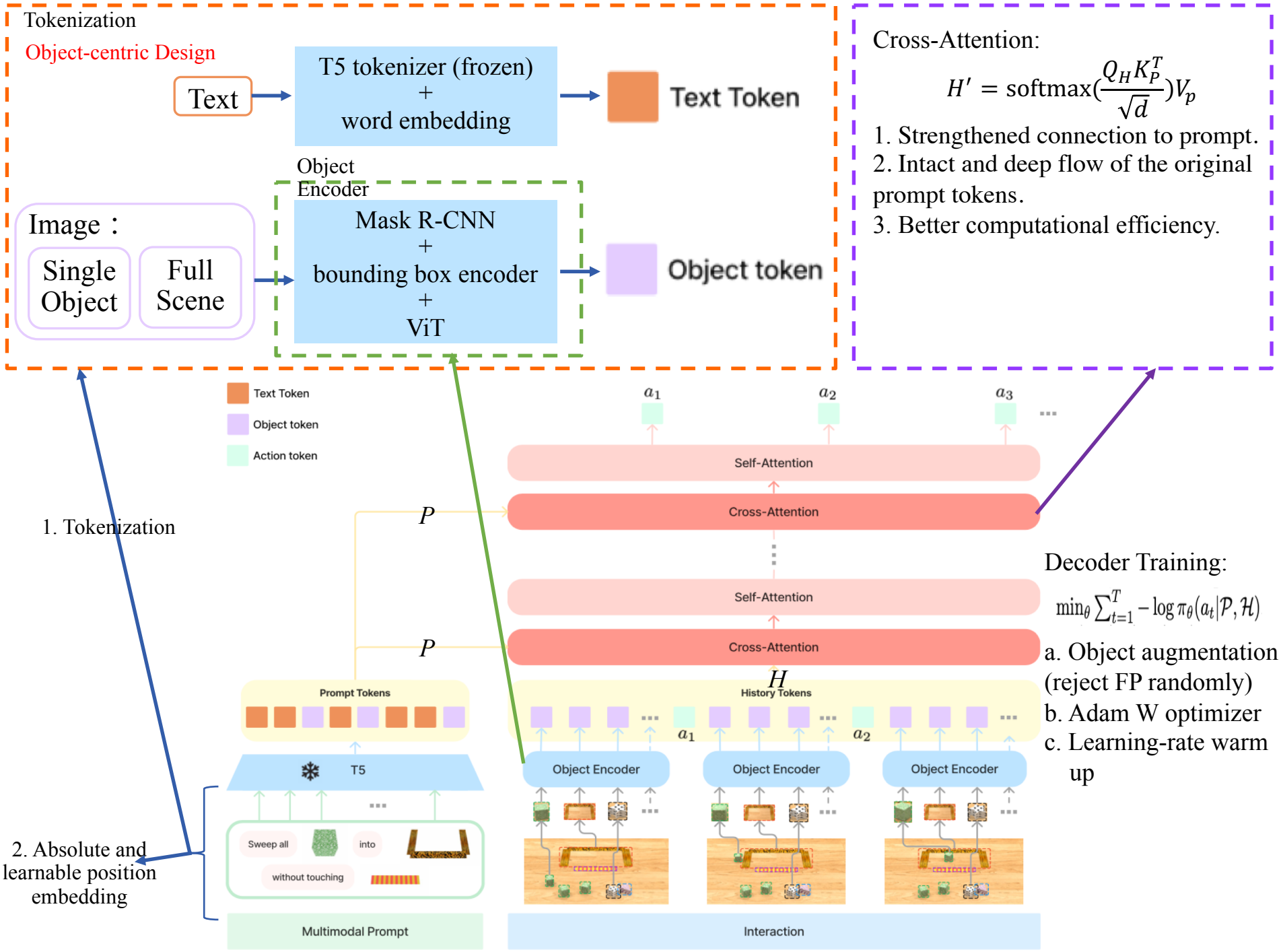
(Reference: Jiang, Yunfan, et al. "Vima: General robot manipulation with multimodal prompts." arXiv preprint arXiv:2210.03094 (2022).)

Problem: Generalist Robot Agent for A Wide Spectrum Task Specifications

Algorithm Architecture: VIMA – Minimalistic Multi-task Encoder-Decoder Architecture with Object-centric design.
Encoder: Generate prompt tokens.
Decoder: Generate actions from history objects condition on prompts.

Benchmark:




Tokenization
Object-centric Design
Text
T5 tokenizer (frozen)
+
word embedding
Text Token
Object
Encoder
Image :
Single Object
Full Scene
Mask R-CNN
+
bounding box encoder
+
ViT
Object token
Cross-Attention:
$H' = \text{softmax}(\frac{Q_H K_P^T}{\sqrt{d}})V_p$
1. Strengthened connection to prompt.
2. Intact and deep flow of the original prompt tokens.
3. Better computational efficiency.
Text Token
Object token
Action token
1. Tokenization
2. Absolute and learnable position embedding
Prompt Tokens
T5
Sweep all
into
without touching
Multimodal Prompt
P
P
H
a_1
a_2
a_3
Self-Attention
Cross-Attention
Self-Attention
Cross-Attention
History Tokens
Object Encoder
Object Encoder
Object Encoder
Interaction
Decoder Training:
$\min_\theta \sum_{t=1}^{T} -\log \pi_\theta(a_t|\mathcal{P},\mathcal{H})$
a. Object augmentation (reject FP randomly)
b. Adam W optimizer
c. Learning-rate warm up

Summarization and Personal Thinking

Summarization of innovations:

1. Prompt-based Generalist Robot Agent for A Wide Spectrum Task Specification:
   a. Generalization: Frozen T5 as encoder for generating prompt tokens with multimodal inputs (interleaving textual and visual inputs).
   b. Task Specification: Decoder consists of alternating cross-attention and self-attention layers for specific tasks learning.
   c. Cross-Attention: History tokens attend to prompt tokens to condition history on prompt.
2. Encoder-Decoder Architecture based Generalist Robot Agent
3. Objects as Tokens
4. Object-centric design (by object detection) for more accurate object-based manipulation learning.

Personal thinking:

1. Since all experiments are simulated, noise sampling and randomness modeling should be considered when applying this architecture in real-world scenes.
2. Upgrade encoder to multitask learning based architecture may achieve better generalization since the variety of robot task types in reality.