Proposal for End-to-End Language + Image to Joint Control

1. Unified Attention Architecture: This proposal is based on a modified attention-driven structure reminiscent of transformers. Our encoder seamlessly integrates image data (augmented with depth information) and textual input. Initial image processing is facilitated via a Convolutional Neural Network, leading to a subsequent analysis through a Vision Transformer.

2. Dual Decoder System:

- **Shared Encoder Motivation**: A shared encoder for both decoders is a tactical choice. By synergizing the loss functions of both decoders, the encoder refines its focus and directionality. This method ensures that encoded details are adept at predicting joint positions while staying attuned to immediate visual objectives.
- **Keypoint Decoder (Inspired by RVT)**: Processes the ViT encoder output to spotlight pivotal image regions. This establishes tangible visual milestones for the model, further bolstering the encoder's alignment with mission-critical elements.
- **Joint Position Decoder (Inspired by ACT)**: Extracts insights from the ViT encoder output, further informed by the Keypoint Decoder, to anticipate joint positions. By integrating ACT's temporal ensemble technique, we counteract the compounding errors intrinsic to imitation learning. Controllers are thus supplied with joint position directives inherently geared towards realizing the visual goals delineated by the Keypoint Decoder.

3. Real-time Memory Integration: To optimize real-time adaptability and nuanced transition understanding, I recommend embedding memory into our transformer architecture. Inspired by the Vision Transformers Need Registers study, I propose discrete memory allocations for each attention head, functioning as registers fed in an autoregressive manner. Such a granular memory approach preserves feature-centric insights, facilitating sequential task comprehension.

4. Language Model Encoder: The current configuration demands unambiguous commands. Anticipating future advancements, a Vision-Language Model (VLM) encoder is a prospective addition, capable of crafting abstract semantic representations. This allows the robot to interpret varied command inputs for the same actions, underlining the necessity for profound research in reasoning and strategic planning.

5. Learning Methodology: While the foundation is set by imitation learning, a dual-faceted strategy blending teleoperation-driven imitation learning with sustained reinforcement learning is advised. Certain studies, such as the Q-Transformer Scalable Offline Reinforcement Learning via Autoregressive Q-Functions, are venturing into analogous terrains.

Note: A version with initial results and implementation details is actively being developed here