**Toward Visual AI Assistants: Enhancing Video Question Answering through Modular and Knowledge-Augmented AI Agents for Real-World Applications**

**Background:** Video Question and Answering (VQA) is a multi-modal artificial intelligence task that combines computer vision and natural language processing. The goal of VQA is to create a system that can understand the contents of a video clip, answer questions about it, and, ultimately, contribute to the development of visual AI assistants while enhancing multi-modal artificial intelligence.

Current state-of-the-art approaches such as Video-LLaMA utilize end-to-end multi-modal neural networks. These models encode both the question and video into latent-space representations using massively pre-trained neural network encoders, and generate an answer using massively pre-trained large language models as a decoder.

This approach performs well on academic benchmarks like ActivityNet or VideoChat, but fails to generalize in applicable real-world scenarios. In my experience applying existing VQA models to build visual assistants at Armada AI, these models capture broad concepts in the videos but lose fine-grain details during the video encoding process. They struggle with tasks such as counting objects or recalling positional relations of objects.

Current approaches encode the video context poorly and lose rich details when obtaining a latent-space representation for video frames that aligns to a static latent-space representation for language. This effectively creates a "visual sentence" for a set of video frames. Videos and natural language have inherently different modalities: Videos are visual and dynamic, while language is symbolic and static. When videos are forced into the same latent space as language, there is a loss of valuable visual information. The complexity of the video content may be compressed or misrepresented in this shared space.

Moreover, current VQA systems fall short due to their inability to readily integrate knowledge without fine-tuning. In practical scenarios, one often requires the inclusion of domain-specific information to answer questions effectively. For instance, a current VQA system would struggle to make sense of a question like "Did Jason's car depart from the parking lot after 5 PM?" Such systems would also fail to responding to questions about novel objects. Consider a VQA system tailored for answering questions about a car company's prototyping day for example-- it would be unable to identify various car types and names.

These models' understanding is limited by their pre-trained encoders and large language model decoders. Such limitations pose significant challenges to building real-world applications with VQA systems, especially for applications involving custom objects and additional context.

The current state of evaluation datasets for VQA systems often inadequately assesses performance in challenging real-world scenarios. For instance, in the widely-used ActivityNet benchmark, questions related to temporal relations, spatial relations, and counting account for only 28.4% of the dataset. Instead, a significant proportion of questions are simple yes/no inquiries. Additionally, these datasets typically lack questions that involve objects or concepts beyond common knowledge. Some training datasets even incentivize language models to provide detailed descriptions of scenes instead of concise answers, which, coupled with the absence of negative samples, heightens the risk of hallucinations during inference time—where the model generates incorrect, nonsensical, or not real responses.

**Proposal:** The end-to-end multi-modal model approach to building VQA systems is not appropriate for real-world uses. Approaches to align videos and language are not mature enough yet to build reliable visual AI assistants that generalize to real scenarios.

I propose an alternative approach to building a VQA system using modular large language model agents that focus on improving generalizability, reducing hallucinations, and refrain from relying on parameter fine-tuning to add knowledge. My proposal is based on recent advancements in large language model agents and leverages highly performant models within computer vision.

**Goal 1, Large Language Model Agent for General VQA:** Large language models are able to effectively act as orchestrators of sub-skills in order to solve complex AI tasks. They are able to use sub-skills as tools to plan high-level and long-horizon tasks given in natural language. While large language

models are unable to understand modalities beyond language, they can coordinate a sequence of interactions with specialized skills, which produce textual summaries of their inference on images and video. In the context of VQA, sub-skills could be specific object detection models, versatile open-vocabulary object detection models, object tracking libraries, object counting libraries, and more. This approach effectively bypasses the alignment issues encountered in end-to-end VQA systems, maintaining all relevant information encoded in text.

To achieve a versatile agent capable of VQA, I propose establishing a foundation of fundamental skills, and subsequently introducing specific models for unique objects as needed. This strategy capitalizes on the advantages offered by a modular skill framework. In cases where the agent lacks essential functionalities, an iterative approach will be taken to integrate additional components until the system can proficiently handle a wide range of sensible VQA tasks.

New findings in large language model alignment research by Meta suggest minimal data is required to align large language models to specific tasks. Many publicly available models are fine-tuned to excel at conversational tasks. I propose aligning them to VQA to improve accuracy.

Goal 1 could be tested on established benchmarks. I will need to design additional benchmarks while assessing the practicality and effectiveness of incorporating new skills.

I anticipate that this approach may yield subpar results in certain VQA tasks compared to the current state-of-the-art approach, particularly those involving conversational reasoning about videos, such as questions like "Why are the kids standing up?" or "Why do people typically enjoy playing this sport?". However, I view this as a necessary trade-off for creating a system capable of broader object generalization and capable of real-world use.

**Goal 2, Enhancing Domain-Specific VQA:** Large language model decoders have two main gaps in their knowledge: (1) the evolving nature of information since their pre-training and (2) the accessibility of specific data on the internet. Consider a large language model agent from Goal 1 attempting to answer the question "Who owns that red car in the parking lot?". The agent might be able to verify the red car exists but would lack knowledge of its owner.

One method to address this issue in purely text-based problems is Retrieval Augmented Generation (RAG)--which leverages prior documents or other verifiable sources of data to fill in the knowledge gaps. Typically, this is done by encoding text paragraphs from such data sources into a latent space and querying the most relevant paragraphs in response to a question. During inference, textual data is queried and supplied alongside the question as context.

To tackle domain-specific inquiries regarding visual content, we need a means of connecting visual data within videos to external knowledge bases. For instance, if we had images of the car linked to a database, we could have answered the question "Who owns that red car in the parking lot?" RAG consists of two essential components: embedding and retrieval.

Developing robust image embeddings presents challenges, particularly when considering real-world applications. A substantial effort will be devoted to crafting embeddings that remain invariant to factors such as image translation, rotation, object orientation, and lighting conditions.

Retrieving documents based on image embeddings of video also poses challenges and is contingent upon the quality of the embedding. Aligning images with relevant documents, especially in scenarios with potentially low-quality images like surveillance footage, is expected to be a complex task.

**Intellectual Merit:** This proposal explores a modular approach to multimodality, assessing the feasibility of large language model agents in addressing real-world problems such as VQA. It aims to understand alternative approaches to multimodality and will contribute insights into optimal approaches for the development of highly adaptable systems in the future.

**Broader Impacts:** This effort has the potential to pave the way for the development of Visual AI agents capable of real-time video comprehension, advancing applications in robotics and embodied assistants. This can particularly aid visually impaired and disabled individuals who aren't able to reason about video on their own. Furthermore, it can provide effective solutions for managing large volumes of live and pre-recorded video streams simultaneously.