2026 START Program
CFP Brief

THEME: **01.** **Robotics/Physical AI**

SUB-THEME: 1.1. Unifying 3D understanding, memory, and prediction

Context/ Overview

Computer vision and 3D perception have made rapid advancements in open-world perception recent years due to intense interest in multi-modal foundation models, NeRFs, Gaussian Splats, video generation, image-to-3D, dense tracking, and related techniques. However, choosing the right perception pipeline for robotic applications involving navigation, manipulation, and inspection is still a matter of trial and error and task-specific engineering. Moreover, image-based pipelines are typically decoupled from useful spatio-temporal functionality such as object pose tracking, spatial memory, agent trajectory prediction, and video Q&A. Samsung is interested in advancing the perceptual capabilities of future robotics applications, such as humanoids and service robots. This project envisions that progress in robotics will be accelerated via human-like perception systems that provide spatial understanding, memory, and prediction in a unified query framework. Outside of robotics, unified 3D perception could also make an impact on XR, wearables, and home IoT devices.

Problem Statement

This project asks to advance methods for unified spatio-temporal perception systems that provide diverse capabilities for recognizing objects, retrieving events from memory, and predictions about the future. Unification of these capabilities is needed to avoid fragmentation "reinventing the wheel" for robot perception systems. The system should accept video input (at a minimum) from a moving camera and allow answering queries about objects, space, and past or future events/trajectories/distributions.

Objectives & Scope

The system may include a database-like memory that can answer queries that may involve text ("how long has my owner been watching TV?"), images (retrieve the mesh of the object that looks like this picture), and/or geometric-temporal queries (yield all objects seen within 0.5m of the table bounding box over the last month). It should also be able to predict aspects of the objects' futures, such as short-term motion or distributions of future locations, or video predictions in a similar manner to world models.

Preferred evaluation settings would include tracking objects, furniture, and people in cluttered indoor environments. Important performance considerations include accuracy (with greater emphasis near the current point in time), real-time updates, query responsiveness, and memory usage. Techniques that operate either entirely on low-power devices or via thin edge clients would be preferred but not necessary.

Successful proposals will outline an innovative, technically sound, and high-impact multi-year effort that builds methods, software systems, and/or benchmarks for understanding environments with thousands of diverse objects.

Specific Topics & focus areas*

Addressing the proposed problem may include research in the following areas:

1. Multi-modal transformers
2. Open-world 3D reconstruction
3. Video Q&A
4. World models
5. On-device 3D & semantic perception

※ The topics are not limited to the above examples and the participants are encouraged to propose other original ideas.

END OF DOCUMENT