# CacheInf: Collaborative Edge-Cloud Cache System for Efficient Robotic Visual Model Inference


## Abstract

Placeholder Placeholder Placeholder Placeholder


## 1 Introduction

Visual information is vital for various robotic tasks. Fast robotic visual model inference is important. While edge devices (robots) typically are limited in computation power, naively offloading the computation is also prone to the limited bandwidth of wireless networks.

We observe that the robotic visual model mostly leverages local operators (i.e., computation results relies on local geometry), and the images for robotic visual model inference are typically continual, meaning that part of the previous computed results of local operators can be cached and reused. This provides opportunity to both reduce data volume to be transmitted while offloading and reduce computation time on both the robot and the server.

We propose CacheInf... Schedule cache and offloading... Reduce data volume transmission in offloading at higher bandwidth... At low bandwidth where offloading is not feasible, reduce local computation time...

The first challenge is to transform cached results to operator acceleration. We reimplemented convolution operators...

The second challenge is how to properly manage the cache memory consumption, especially at the robot side which typically has a tight GPU memory budget.

The third challenge is since cache both exists on the robot and the server, how to ultimately leverage the existing cache to accelerate inference. Introduce Intra-DP (TODO: cite arxiv name) to parallelize local cached computation and transmission...

We implemented CacheInf...

Evaluation shows that...

The major contribution of this paper is ...

## 2 Background

## 3 System Overview

## 4 Implementation

## 5 Evaluation

## 6 Conclusion

## 7 Conclusions

## References