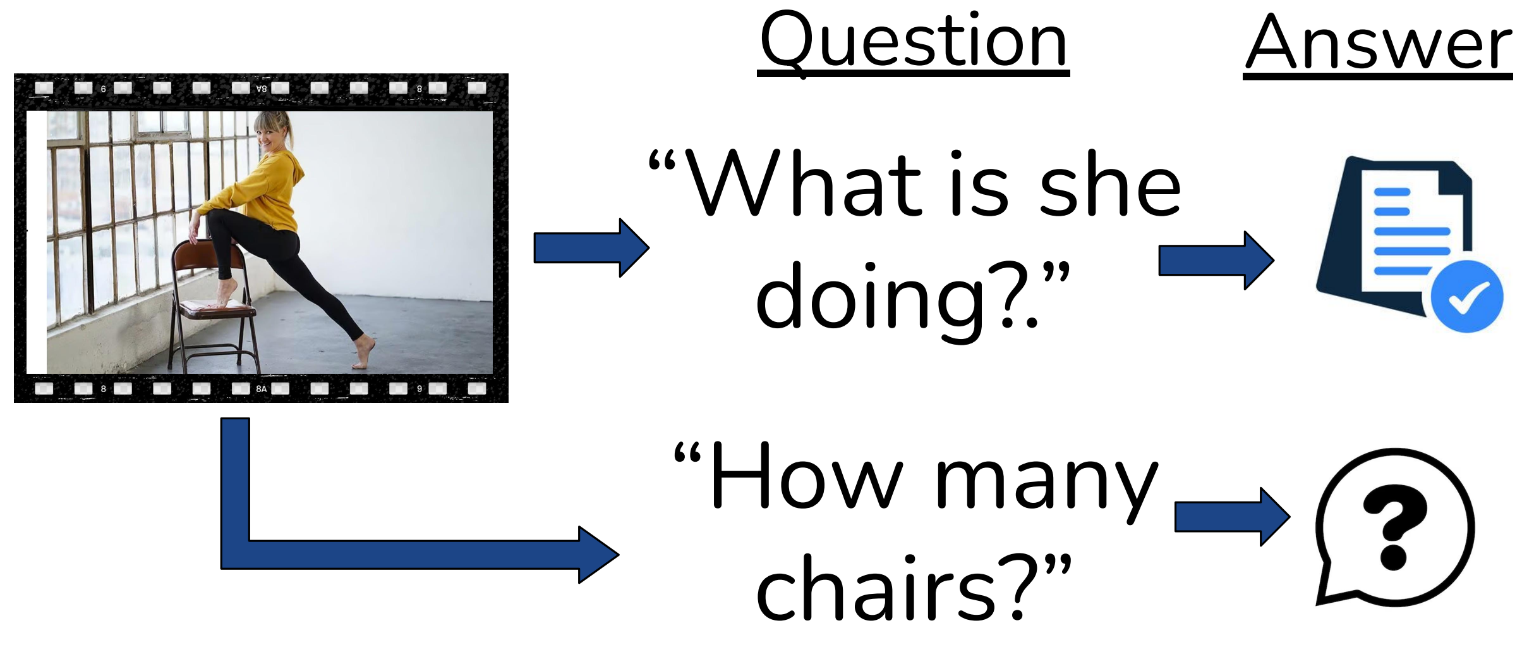
VisCount: Evaluating Counting Capabilities of Multimodal Video-LLMs

Swetha Krishnan Rahasya Barkur Kushal Raju

UMassAmherst | Manning College of Information & Computer Sciences

Overview

Video Understanding with Video-LLMs



Question Answer

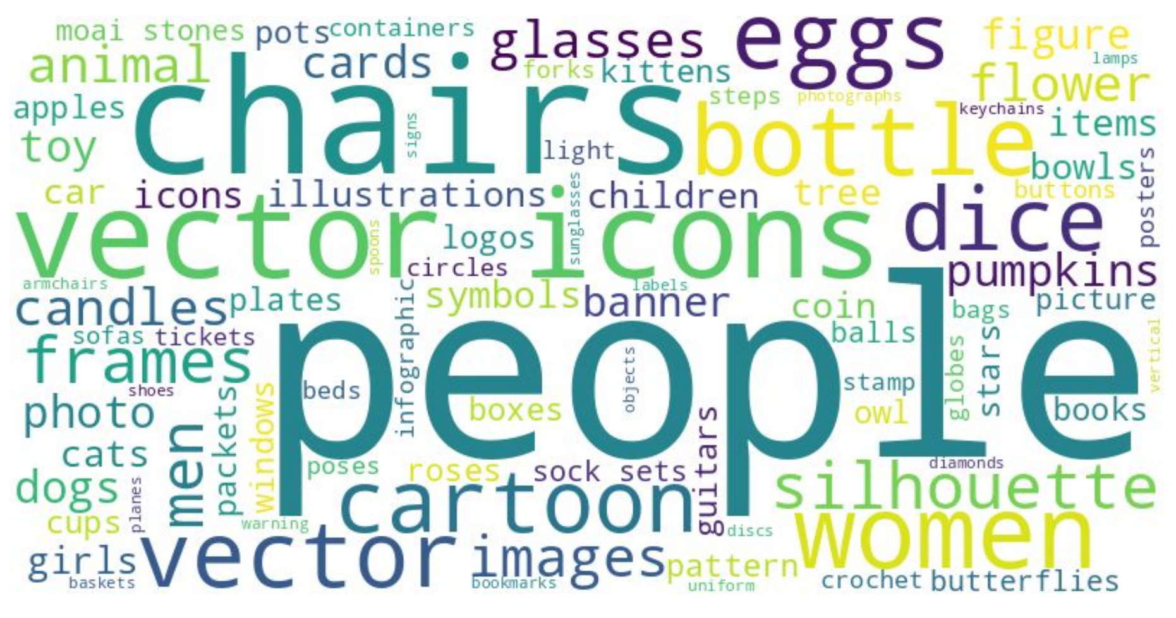
"What is she doing?."

"How many chairs?"

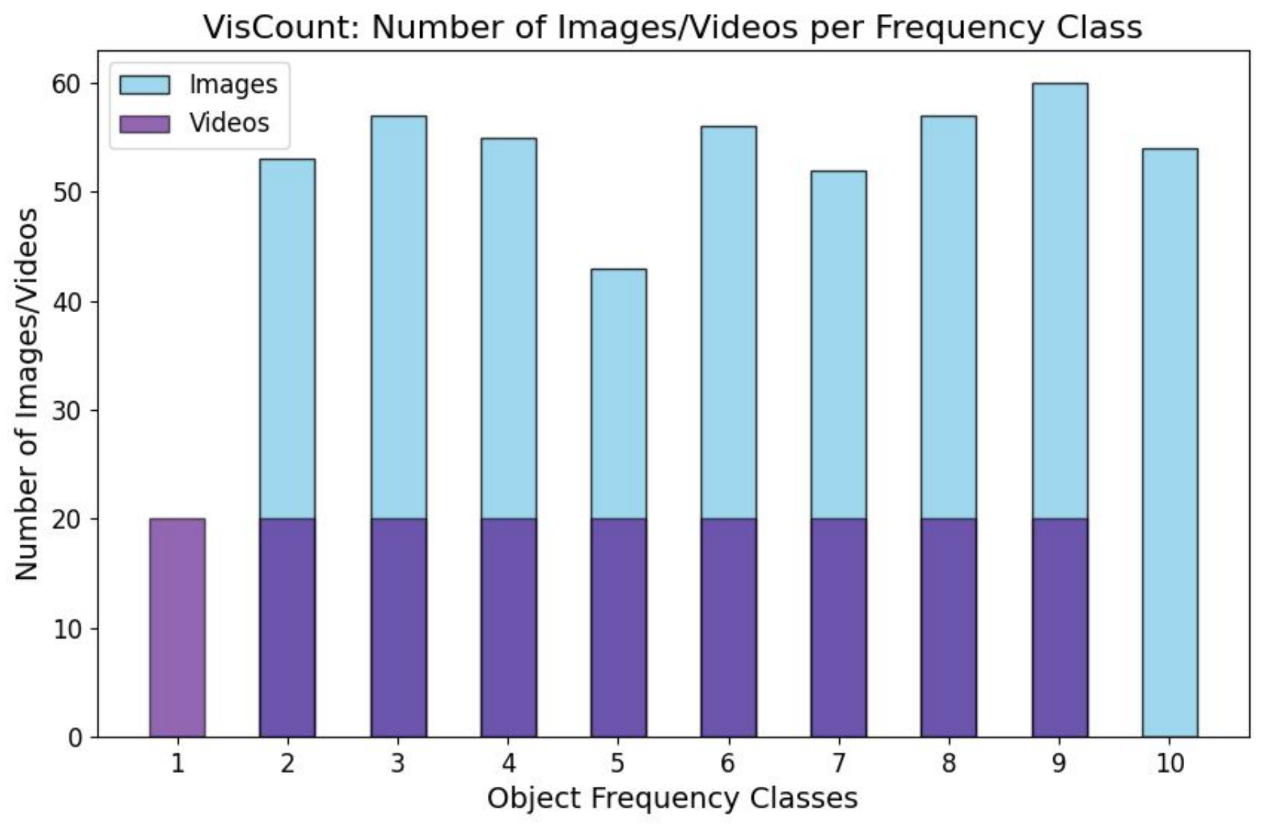
Can multimodal Video-LLMs count accurately?

- **Goal:** generate a vision dataset which:
  - evaluates object counting,
  - spans images and videos,
  - includes low-density, visible and widely recognised objects

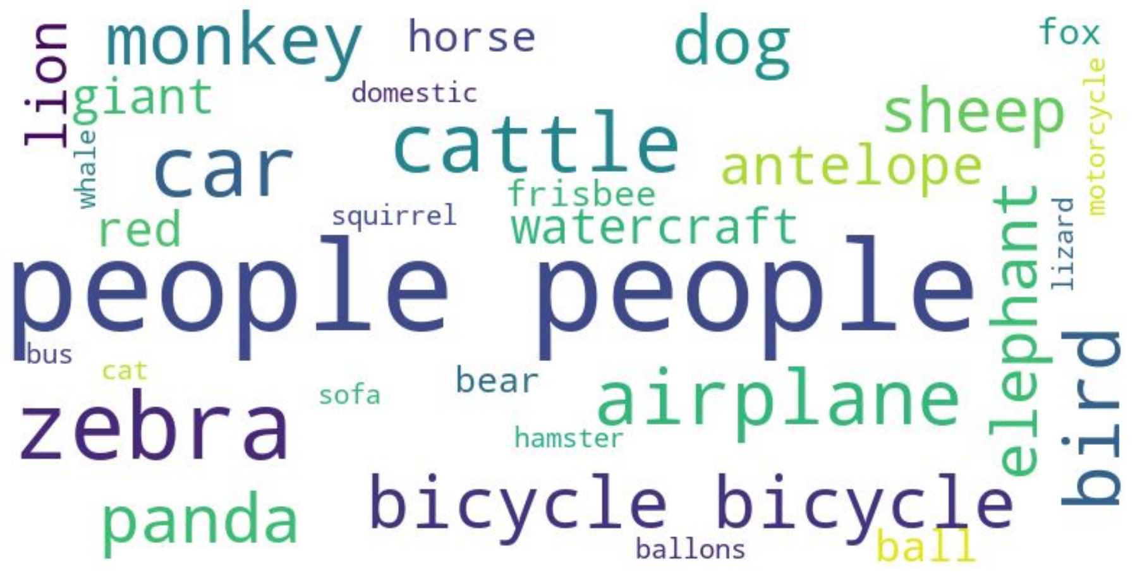
Dataset



Object Categories: VisCount Images

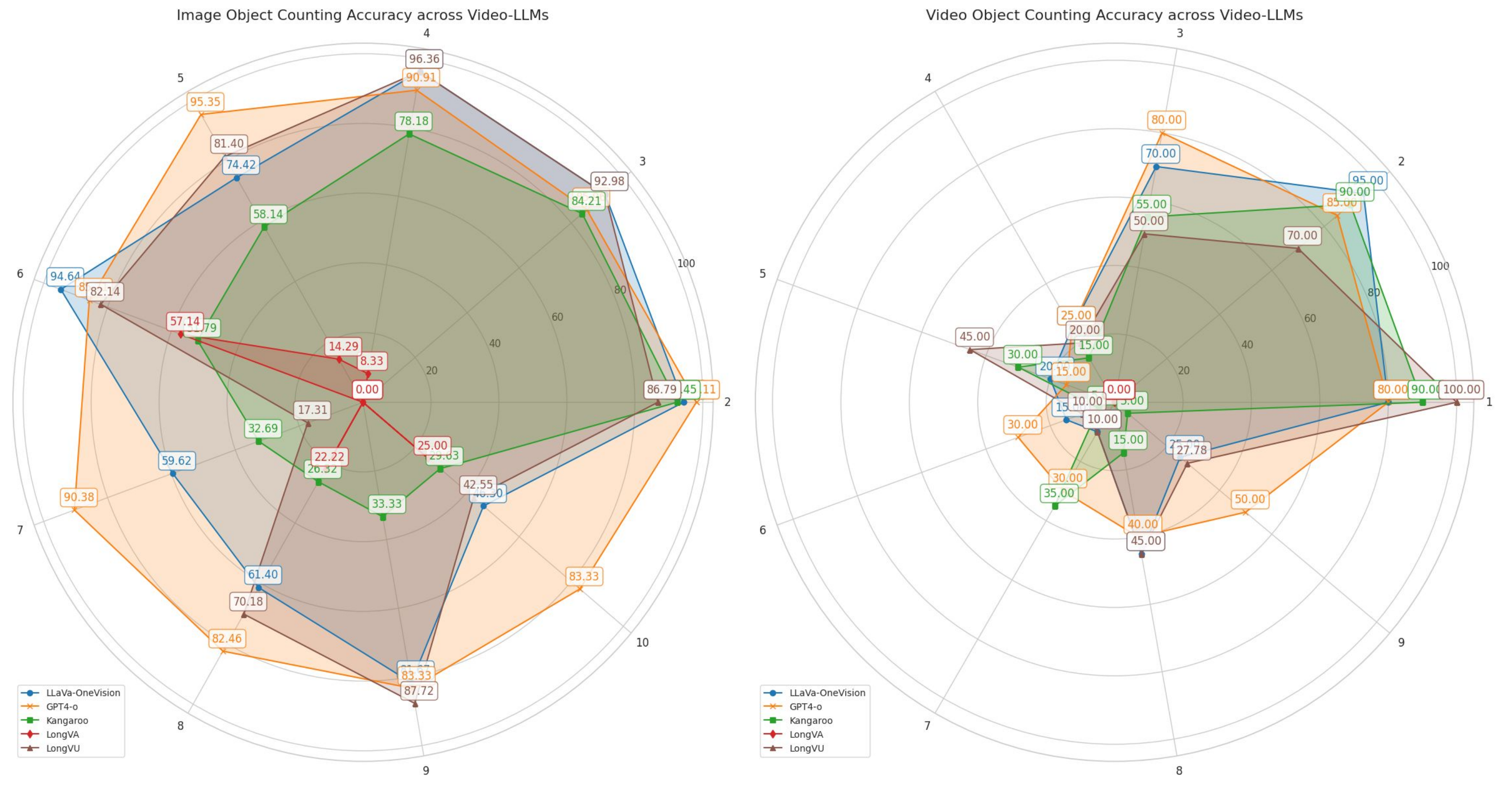


Frequency Distribution



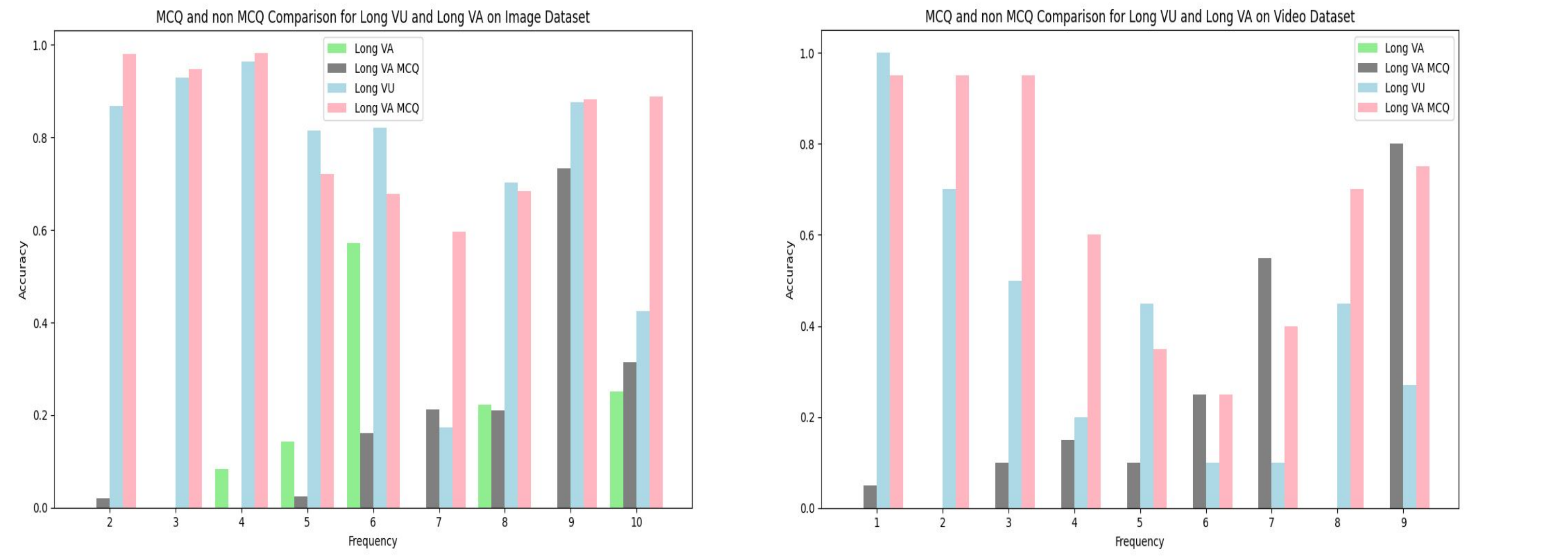
Object Categories: VisCount Videos

Experiments & Results



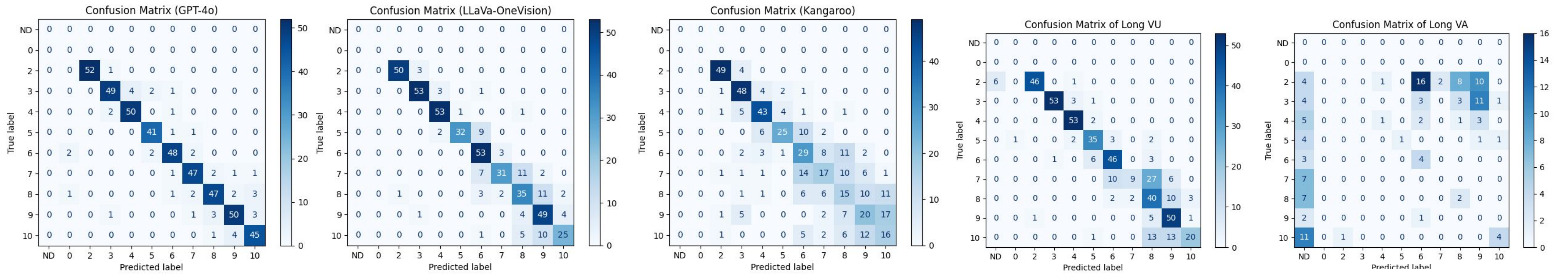
Performance degrades with increasing object frequencies across images and videos

Case Study: Prompting with MCQs

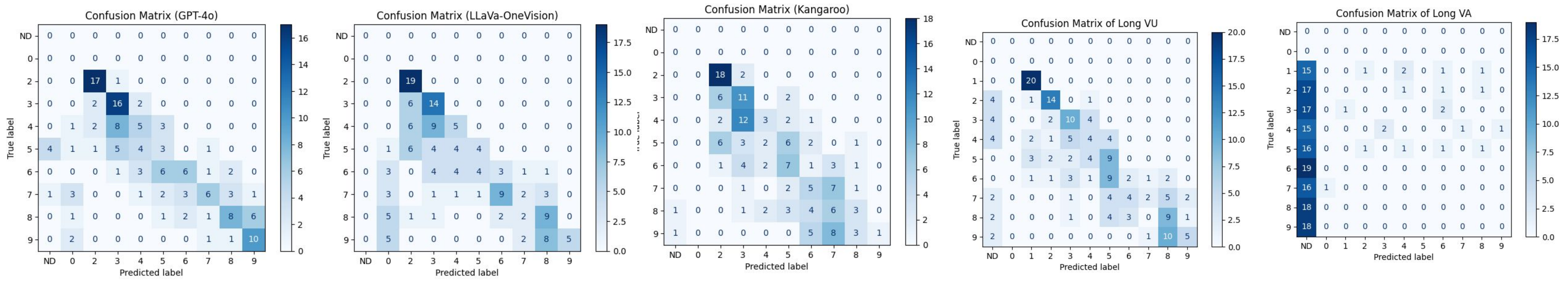


Models count objects better with MCQs

Evaluation



Evaluation on images



Evaluation on videos

Conclusion

- Multimodal video-LLMs are not well-suited for object counting and require improvement in numerical reasoning and object differentiation
- Future steps:
  - Expand dataset with more videos, images and object types
  - Include high object density images and videos