

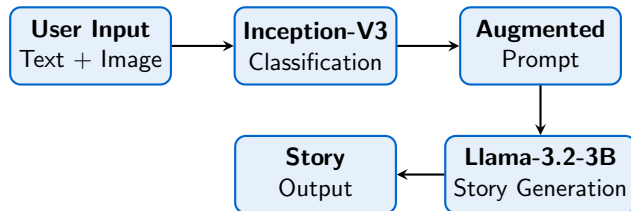
# o!SNAP: Agentic AI Storyteller

## Embedded Systems Workshop - Mid-Eval

Samarth Rao   Neha Prabhu   Avani Sood   Pranshul Shenoy

[github.com/esw-m25-osnap/agentic-ai-storyteller](https://github.com/esw-m25-osnap/agentic-ai-storyteller)

# Our Agentic Pipeline



## Example Results:

- Banana + Water Bottle → Story about thirsty Emma who ate a banana
- Keyboard → Story about Emily working on a project, with focus on her using a keyboard

## Successfully Generated Stories

### Test 1: Banana + Water Bottle

- Both objects identified
- Coherent narrative
- Multiple image support

### Test 2: Keyboard

- Object classified correctly
- Story generated
- External noise Eliminated

*Full stories available in the report appendix*

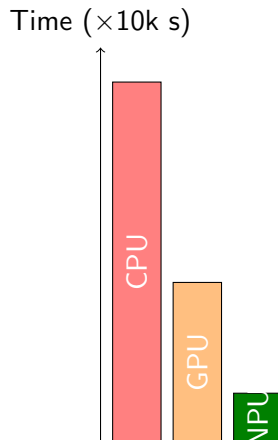
**Let's Look at it in action!**

## Inference Time (Inception-V3):

- CPU: 59,255 s
- GPU: 26,228 s
- NPU: 7,898 s ← 7.5× faster!

## Trade-off: Initialization Cost

- CPU: 0.14s
- GPU: 0.54s
- NPU: 1.43s



*NPU optimal for batch processing, not single inferences*

## Inception-V3

- 1000-class ImageNet model
- Input:  $299 \times 299 \times 3$  images
- Deployed via SNPE
- Runs on CPU, GPU & HTP

## Llama-3.2-3B

- 71 files, 3-part binary
- NLP model with significant context buffer length
- QNN-compatible
- Quantized on NPU

## SNPE

### Snapdragon Neural Processing Engine

- Runs on CPU, GPU, HTP (NPU)
- Optimizes performance & power
- Used for Inception-V3

## Genie

### Qualcomm's GenAI Framework

- Simplifies LLM deployment
- QNN Runtime support
- Used for Llama-3.2-3B

# Takeaways & Next Steps

## What We Achieved


- Deployed Quantized models on QIDK edge device
- Built working agentic pipeline (vision → text → story)
- Benchmarked CPU/GPU/NPU performance
- Validated inference accuracy

## Future Goals

- Extend to **network of edge devices**
- Dynamic agent coordination for model usage
- More sophisticated agentic behaviors
- Collaborative AI workflows



# Thank You!

 [github.com/esw-m25-osnap/agent-ai-storyteller](https://github.com/esw-m25-osnap/agent-ai-storyteller)