ARIA: Optimizing Vision Foundation Model Inference on Heterogeneous Mobile Processors for Augmented Reality

MobiSys 2025 Best Paper Award

马浩然

2025.7.11

章节介绍

- 作者介绍
- 研究背景与动机
- ARIA 系统设计
- 实验评估
- 总结与讨论

章节介绍

- 作者介绍
- 研究背景与动机
- ARIA 系统设计
- 实验评估
- 总结与讨论

作者介绍

Jeho Lee

延世大学 (Yonsei University) 博士生 导师：Hojung Cha 教授

致力于为下一代应用（如移动 AR、自动驾驶系统）构建实时、端侧视觉 AI 系统

研究方向:

- 端侧 AI (On-Device AI)，主攻视觉方向
- 移动与嵌入式系统 (Mobile and Embedded Systems)

论文发表：

- [MobiSys'25] ARIA: Optimizing Vision Foundation Model Inference on Heterogeneous Mobile Processors for Augmented Reality (BestPaper Award)
- [MobiCom'24] Panopticus: Omnidirectional 3D Object Detection on Resource-constrained Edge Devices
- [INFOCOM'24] Vulture: Cross-Device Web Experience with Fine-Grained Graphical User Interface Distribution

章节介绍

- 作者介绍
- **研究背景与动机**
- ARIA 系统设计
- 实验评估
- 总结与讨论

研究背景与动机

AR/VR 发展浪潮

近年来，随着苹果 Vision Pro、Meta Quest 等产品的不断涌现，增强现实（AR）与虚拟现实（VR）技术正以前所未有的速度发展，在工业设计、教育培训和游戏娱乐等领域应用广泛

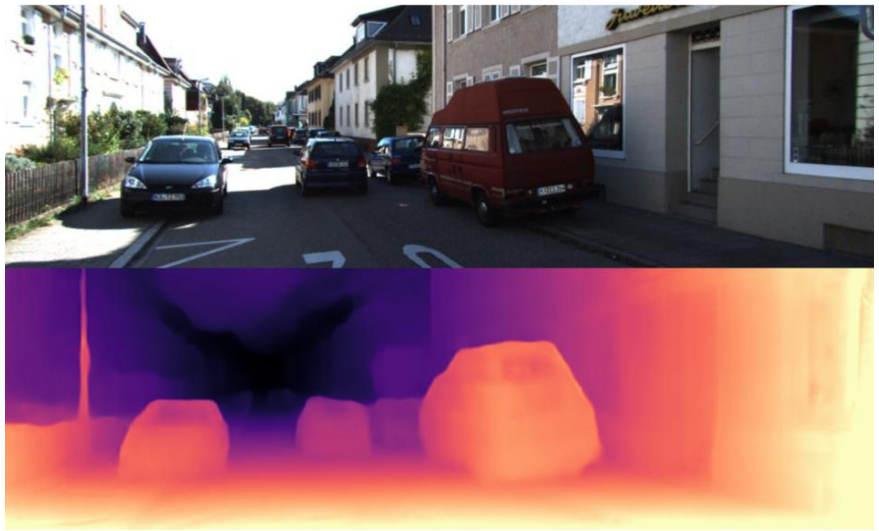
- 波音公司 AR 维修：波音使用 AR 眼镜为技术人员提供实时维修指导
- 医学培训： 医学生通过 AR/VR 头显模拟手术场景，练习复杂手术操作
- Pokémon GO：利用 AR 和 GPS，将虚拟的宝可梦融入现实世界

研究背景与动机

AR 应用对视觉能力的较高需求

移动 AR 应用（比如虚拟物体放置、场景交互）需要对环境进行高质量、实时的视觉理解以支持使用者真实的、沉浸式的体验效果；低准确度的视觉理解或较差的实时性会产生错误或卡顿，破坏用户体验

- 任务案例: 像素级的深度估计（远近关系）和语义分割（物体类别）
- 核心要求: 高精度 + 实时性（例如 30 FPS）



研究背景与动机

现有 AR 神经网络模型技术方案

现有技术方案可以分为

- 传统方案 Domain-Specific Model: 为了实现高精度、实时的视觉理解，业界首先采用了在特定数据集上训练的，参数规模较小的模型（即 DSM）

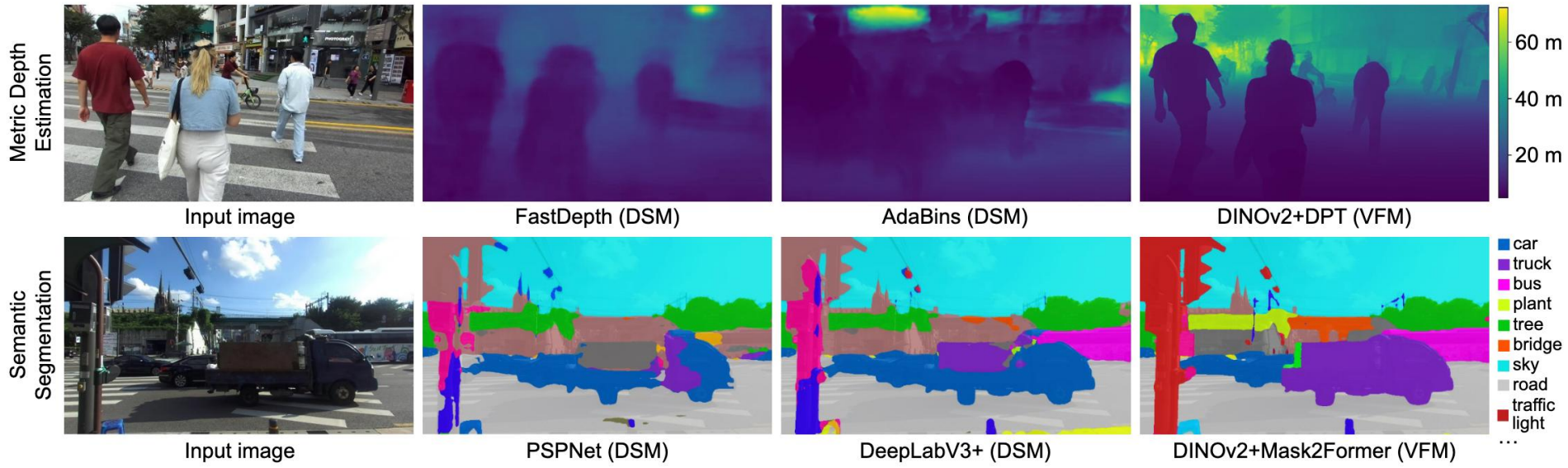
- 新兴方案 Vision Foundation Model: 采用 Transformer 架构，使用大规模数据集进行训练的，参数规模比较大的基础模型（即 VFM）

类型	模型	参数量
DSM	FastDepth（NYUv2）	4.0M
VFM	DINOv2+DPT	24.8M

研究背景与动机

现有 AR 应用模型选择：从 DSM 到 VFM

- 传统 DSM (Domain-Specific Model) 的局限: 在特定数据集上训练，泛化能力差，遇到新场景（如光线变化、新环境）时精度骤降
- VFM (Vision Foundation Model) 的崛起: 基于 Transformer 架构，在海量、多样化的数据上训练 ,具备强大的“零样本”泛化能力，在未见过的数据上依然表现出色



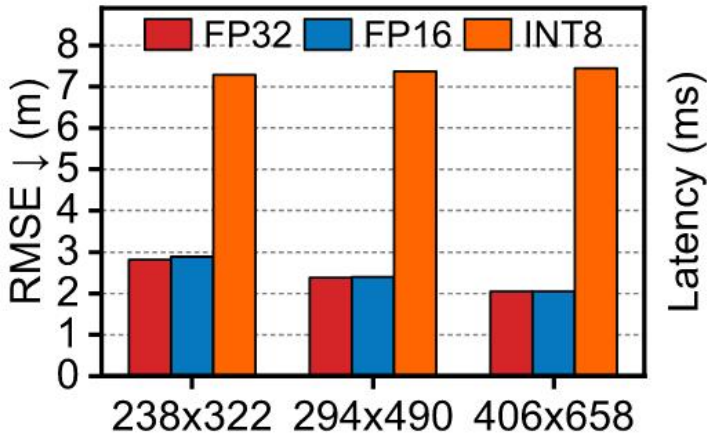
研究背景与动机

VFM 在移动端的“两难困境”

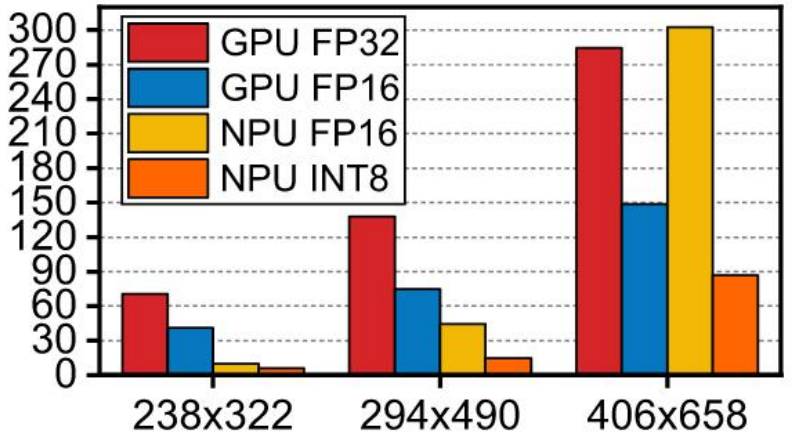
VFM 模型巨大且计算密集，直接部署在移动 AR/VR 设备上非常困难，面临的权衡 (Trade-off):

- 追求高精度 (高保真): 使用高分辨率输入 + FP16/32 浮点运算；结果: 精度很高，但即使在移动 GPU 上运行，延迟也过高，无法满足实时性要求

- 追求实时性 (低保真): 使用低分辨率输入 + INT8 整型量化，在 NPU 上运行速度快；结果: 满足了速度，但由于 VFM 对量化敏感以及信息丢失，导致精度严重下降



(a) Accuracy (H×W)



(b) Latency (H×W)

研究背景与动机

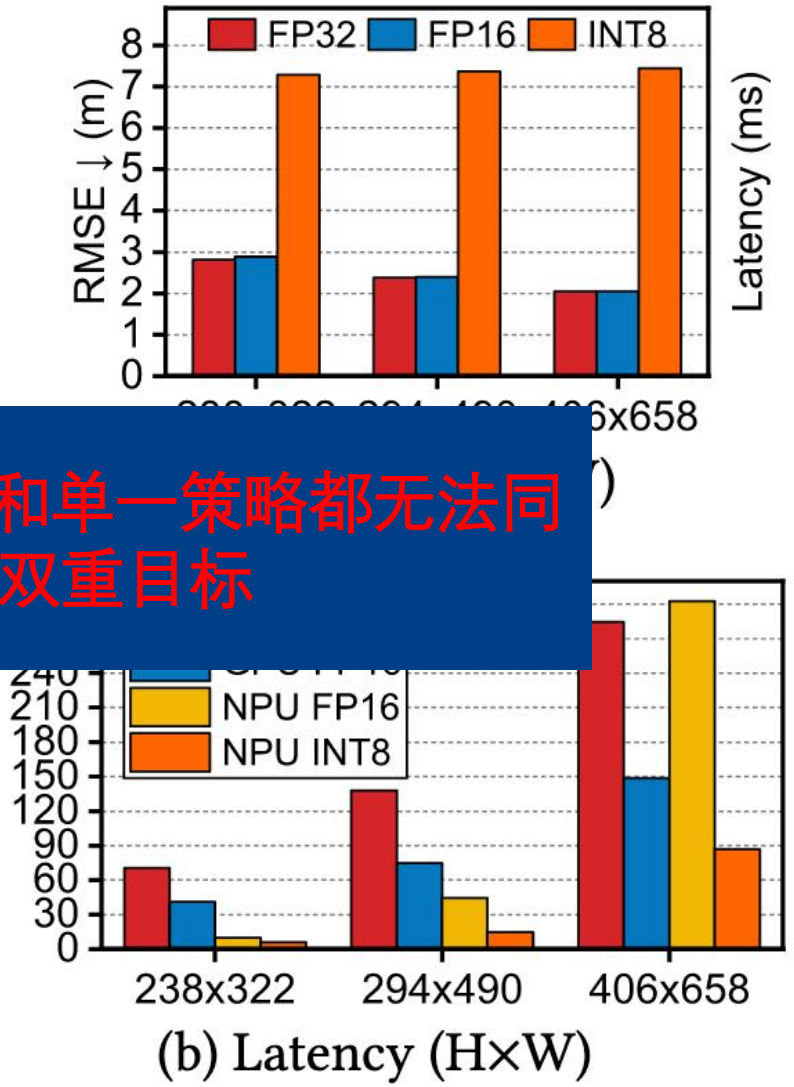
VFM 在移动端的“两难困境”

VFM 模型巨大且计算密集，直接部署在移动 AR/VR 设备上非常困难，面临的权衡 (Trade-off):

- 追求高精
  FP16/32
  在移动 G
  实时性要求

- 追求实时性 (低保真): 使用低分辨率输入 + INT8 整型量化，在 NPU 上运行速度快；结果: 满足了速度，但由于 VFM 对量化敏感以及信息丢失，导致精度严重下降

单一处理器（无论是 GPU 还是 NPU）和单一策略都无法同时满足“高精度”和“低延迟”的双重目标



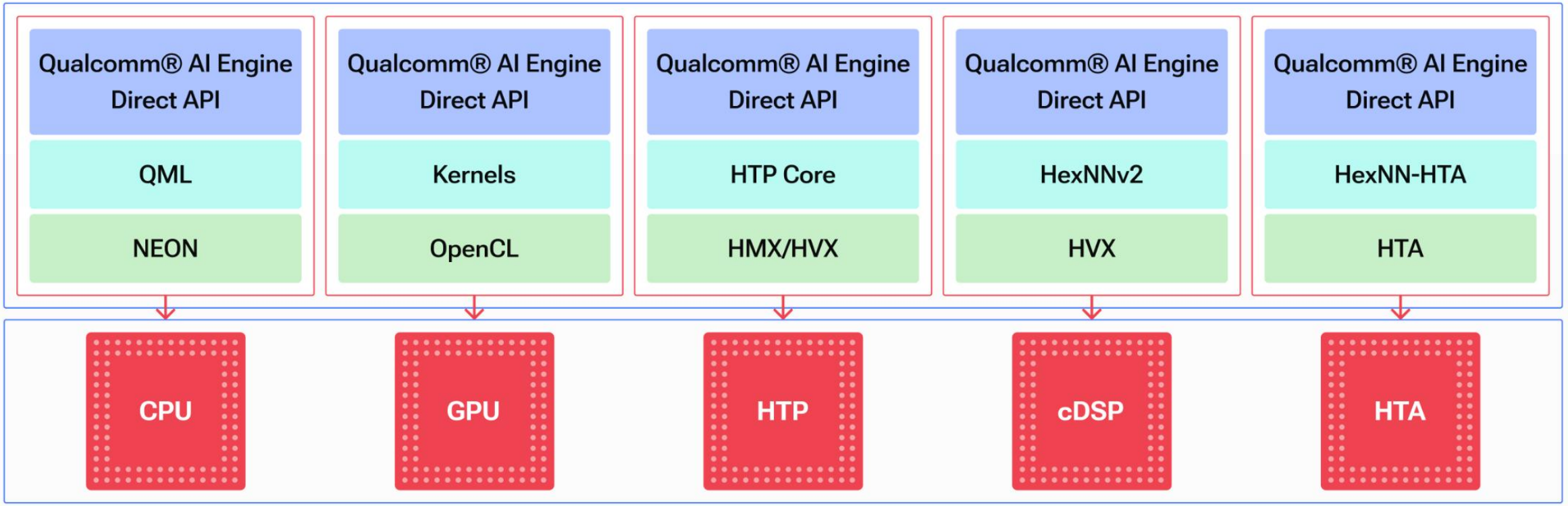
章节介绍

- 作者介绍
- 研究背景与动机
- **ARIA 系统设计**
- 实验评估
- 总结与讨论

ARIA 系统设计

ARIA 的核心思想：并行与选择性推理

打破“二选一”的困境，充分利用移动芯片的异构性 (Heterogeneity)



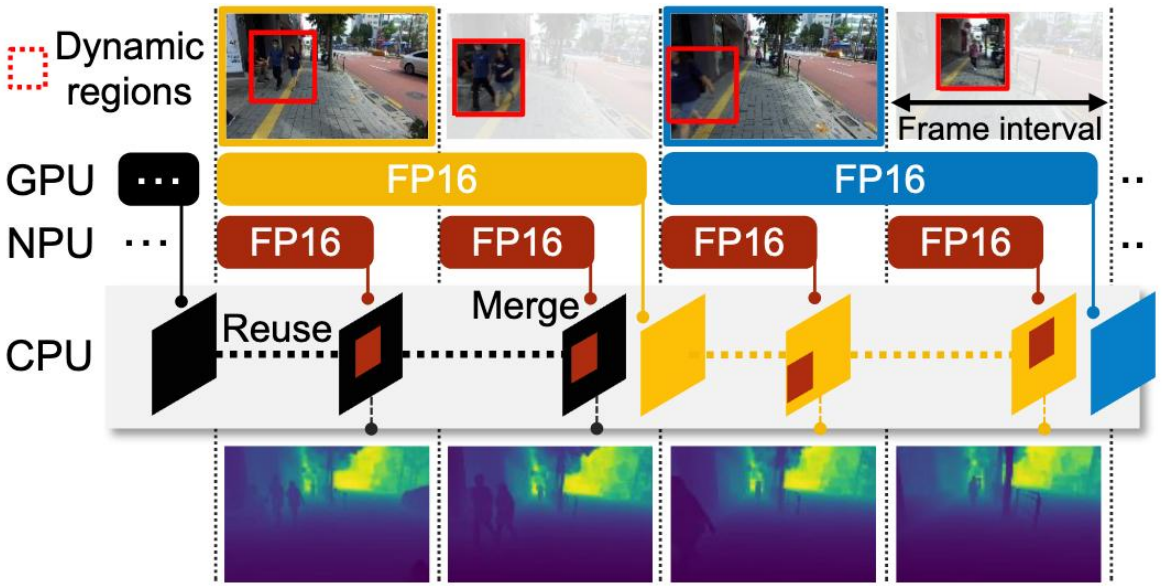
ARIA 系统设计

ARIA 的核心思想：并行与选择性推理

- GPU (高精度): 不追求每一帧都计算，而是周期性地处理完整的、高分辨率的相机画面，生成一个高质量的“背景”特征图

- NPU (低延迟): 在 GPU 计算的间隙，只针对画面中的动态区域（如移动的行人、车辆）进行快速、实时的局部更新

将 GPU 的高质量全局预测和 NPU 的高效率局部更新结合起来，实现高精度低延迟的推理目标



ARIA 系统设计

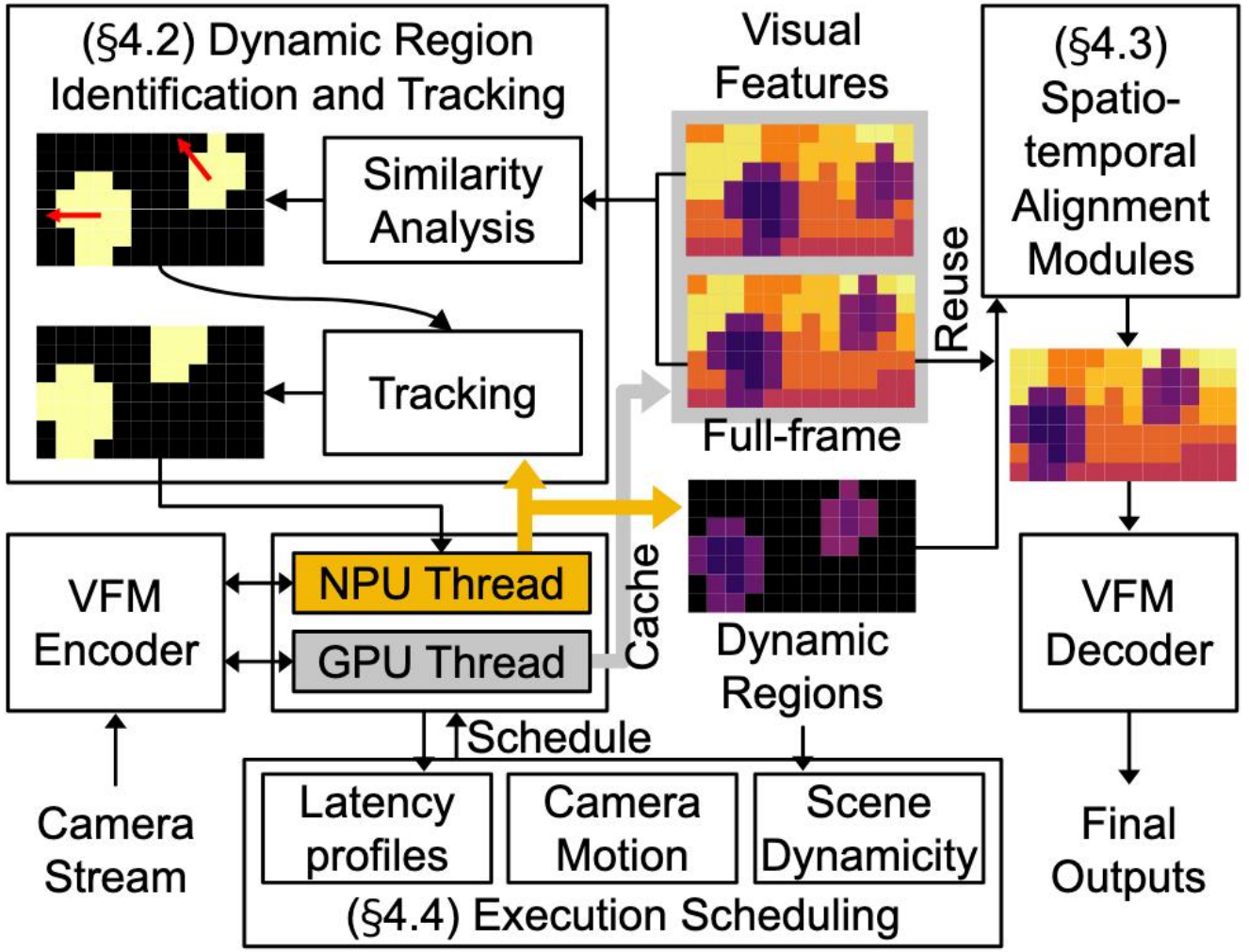
并行与选择性推理面临挑战

- 挑战一：动态区域处理 - 如何高效、准确地识别和处理动态区域
- 挑战二：时空对齐 - 来自 GPU 和 NPU 的“分裂”预测结果如何无缝拼接，避免产生视觉瑕疵
- 挑战三：运行时自适应 - 如何应对设备发热（性能下降）和相机剧烈运动等动态变化

ARIA 系统设计

ARIA 系统架构

- 相机视频流输入后，进入 VFM Encoder
- GPU 线程和 NPU 线程并行工作
- ARIA 的三个**核心模块**分别解决上述三个挑战
- 最终，对齐后的特征图送入 VFM Decoder 生成最终输出

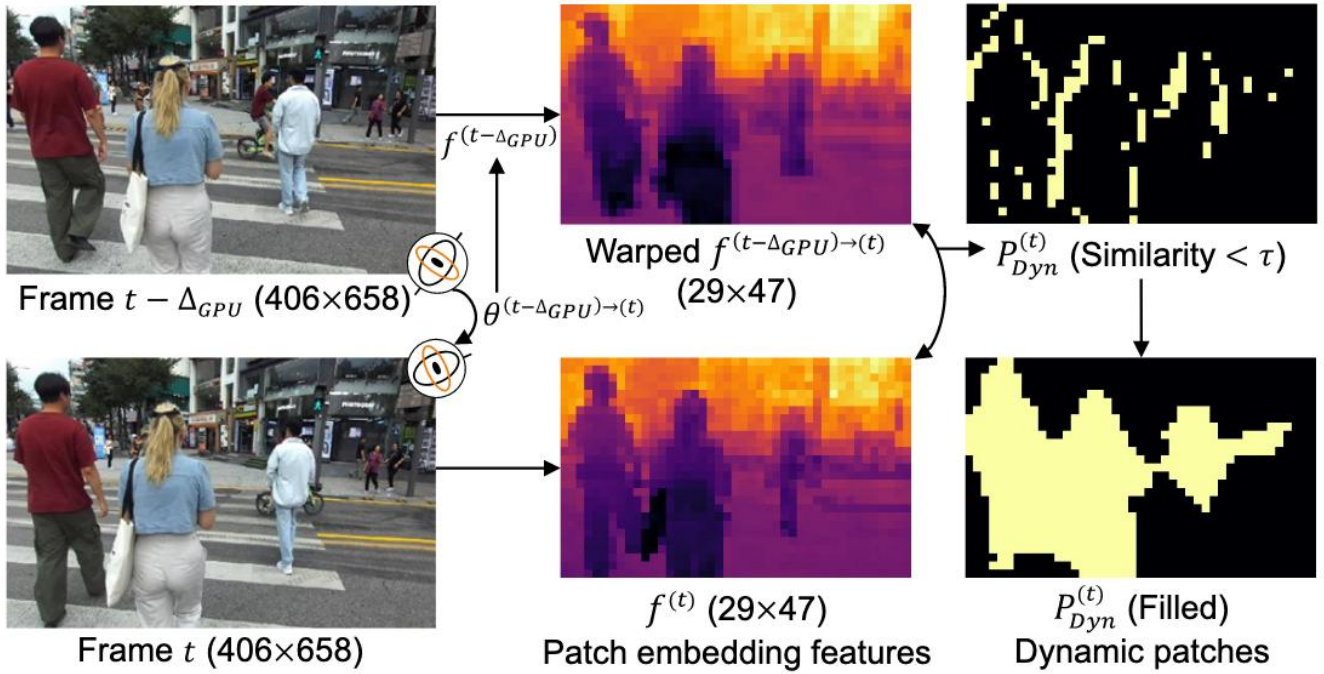


ARIA 系统设计

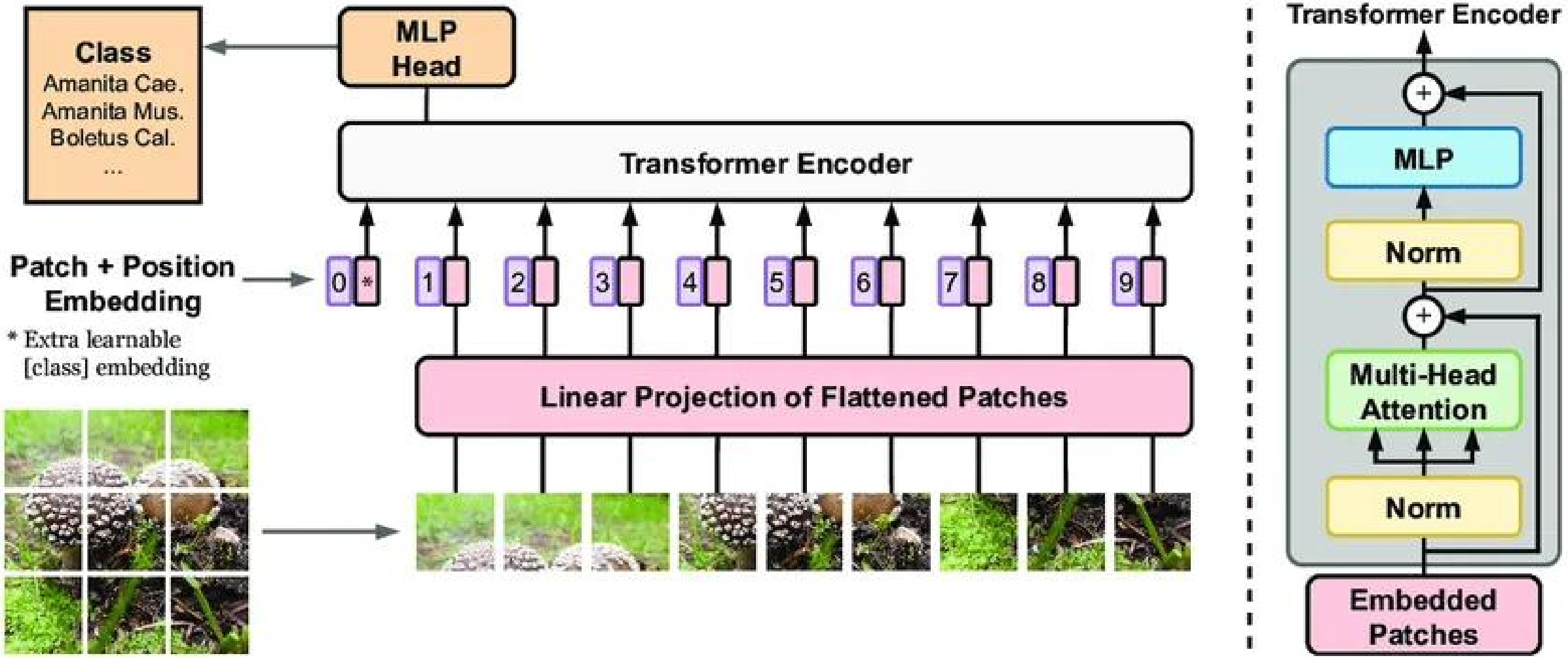
模块一：基于 Transformer 的动态区域识别与追踪

- 识别 (Identification): 利用 VFM Encoder 中间产生的 Patch Embedding 特征，通过对比前后两帧（GPU处理的帧）特征的时域差异（余弦相似度），以 Patch（如 14x14像素）为单位，精细地找出动态区域

- 跟踪 (Tracking): 在 GPU 不工作的帧，我们没法得到全局特征。因此，ARIA 会计算动态区域的运动向量，并在后续帧中对其进行位置跟踪，直到下一次 GPU 更新



ARIA 系统设计

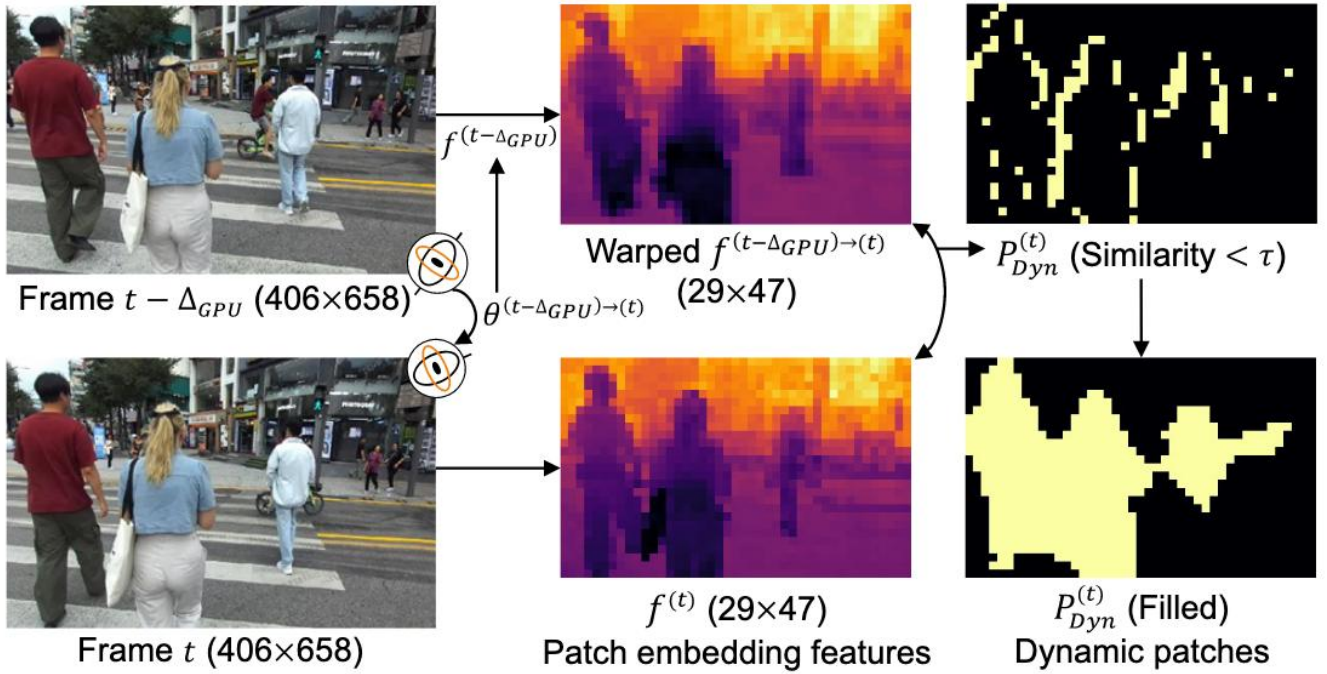

Class
Amanita Cae.
Amanita Mus.
Boletus Cal.
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

ARIA 系统设计

模块一：基于 Transformer 的动态区域识别与追踪

- 识别 (Identification): 利用 VFM Encoder 中间产生的 Patch Embedding 特征，通过对比前后两帧（GPU处理的帧）特征的时域差异（余弦相似度），以 Patch（如 14x14像素）为单位，精细地找出动态区域

- 跟踪 (Tracking): 在 GPU 不工作的帧，我们没法得到全局特征。因此，ARIA 会计算动态区域的运动向量，并在后续帧中对其进行位置跟踪，直到下一次 GPU 更新

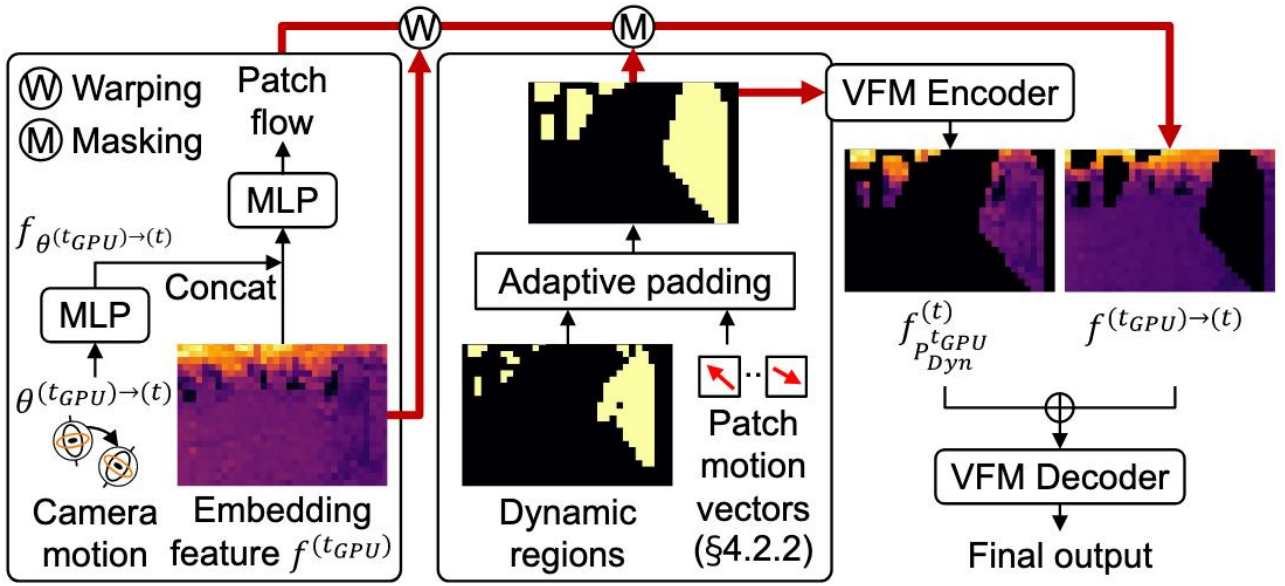


ARIA 系统设计

模块二：感知动态的“时空对齐”模块

- 时间对齐: 当复用上一张 GPU 的结果时，需要将其“变换”到当前视角，ARIA 训练了一个小型 MLP 网络，根据相机运动来预测每个 Patch 的位移，实现更精确的特征图变换，避免视差问题

- 空间对齐: 为了让 NPU 理解动态区域周围的上下文，在处理动态区域时，引入了自适应填充，
即向外扩展一些区域一并处理，
保证了局部和全局特征的平滑过渡 。



ARIA 系统设计

模块三：动态执行调度

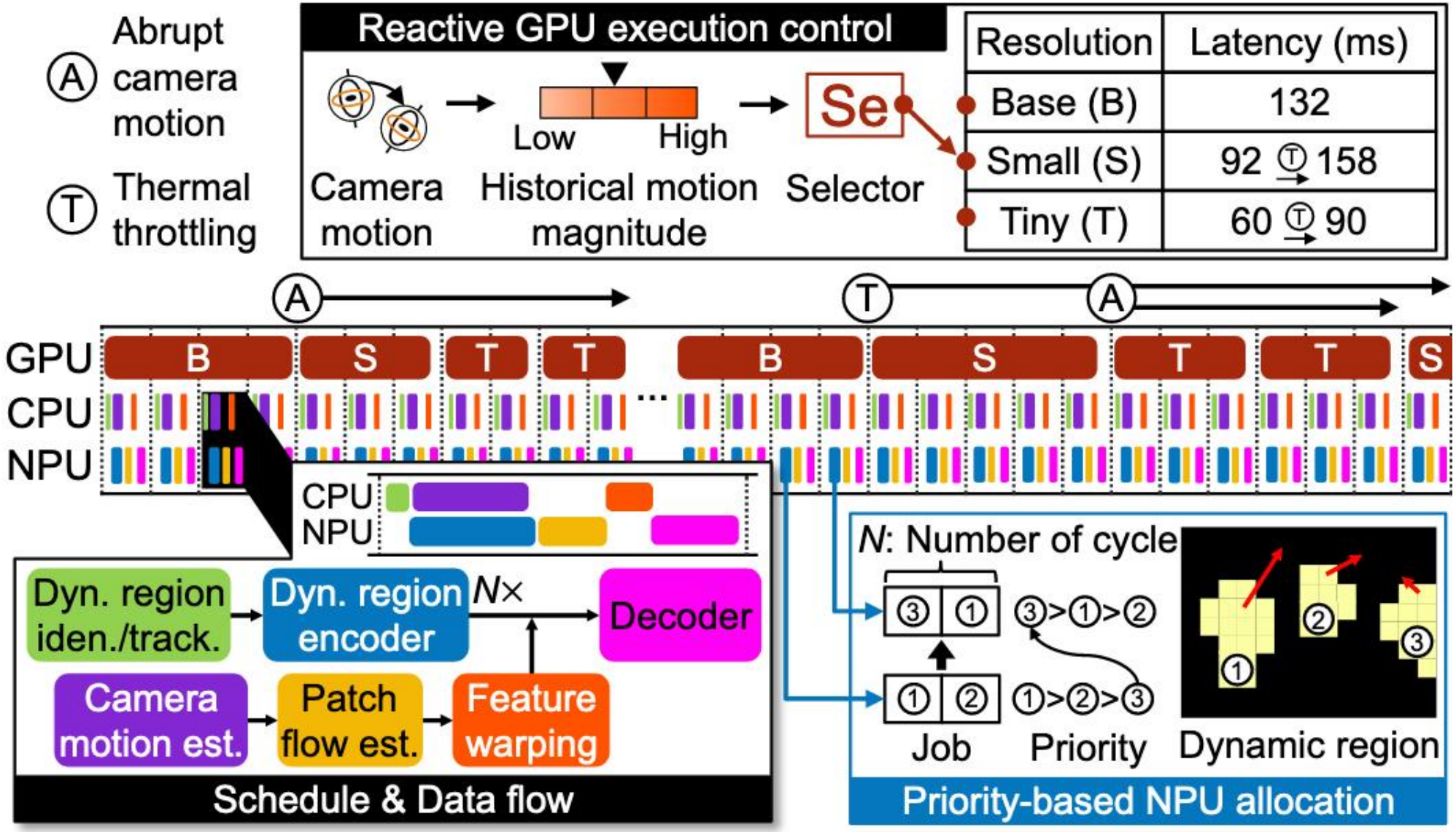
响应式 GPU 调度

- 应对相机剧烈运动: 当检测到相机快速旋转时，旧的 GPU 结果很快就失效了。ARIA 会主动降低 GPU 处理的分辨率，让它更快地生成一张新的全局特征图
- 应对设备发热: 当检测到 GPU 因发热导致处理变慢时，同样会降低分辨率来保证处理速度

基于优先级的 NPU 调度

- NPU 的算力有限，无法在一帧内处理所有动态区域
- ARIA 会根据运动速度给动态区域排优先级，优先更新移动最快的物体，因为这些物体如果延迟，给用户的观感破坏最大

ARIA 系统设计


Ⓐ Abrupt camera motion
Ⓣ Thermal throttling
Reactive GPU execution control
Low
High
Se
Camera motion
Historical motion magnitude
Selector
Resolution
Latency (ms)
Base (B)
132
Small (S)
92 Ⓣ 158
Tiny (T)
60 Ⓣ 90
GPU
CPU
NPU
B
S
T
T
B
S
T
T
S
Ⓐ
Ⓣ
Ⓐ
CPU
NPU
Dyn. region iden./track.
Dyn. region encoder
N×
Decoder
Camera motion est.
Patch flow est.
Feature warping
Schedule & Data flow
N: Number of cycle
③ ①
③>①>②
① ②
①>②>③
Job
Priority
Dynamic region
Priority-based NPU allocation

章节介绍

- 作者介绍
- 研究背景与动机
- ARIA 系统设计
- **实验评估**
- 总结与讨论

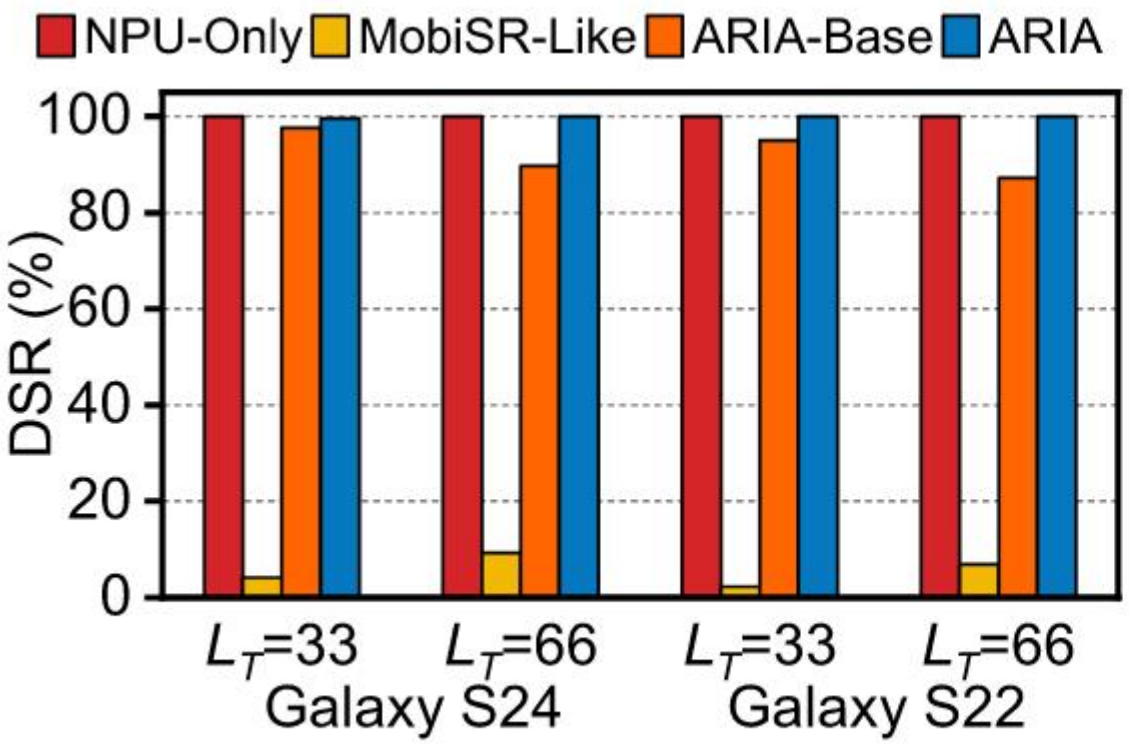
实验评估

实验设置

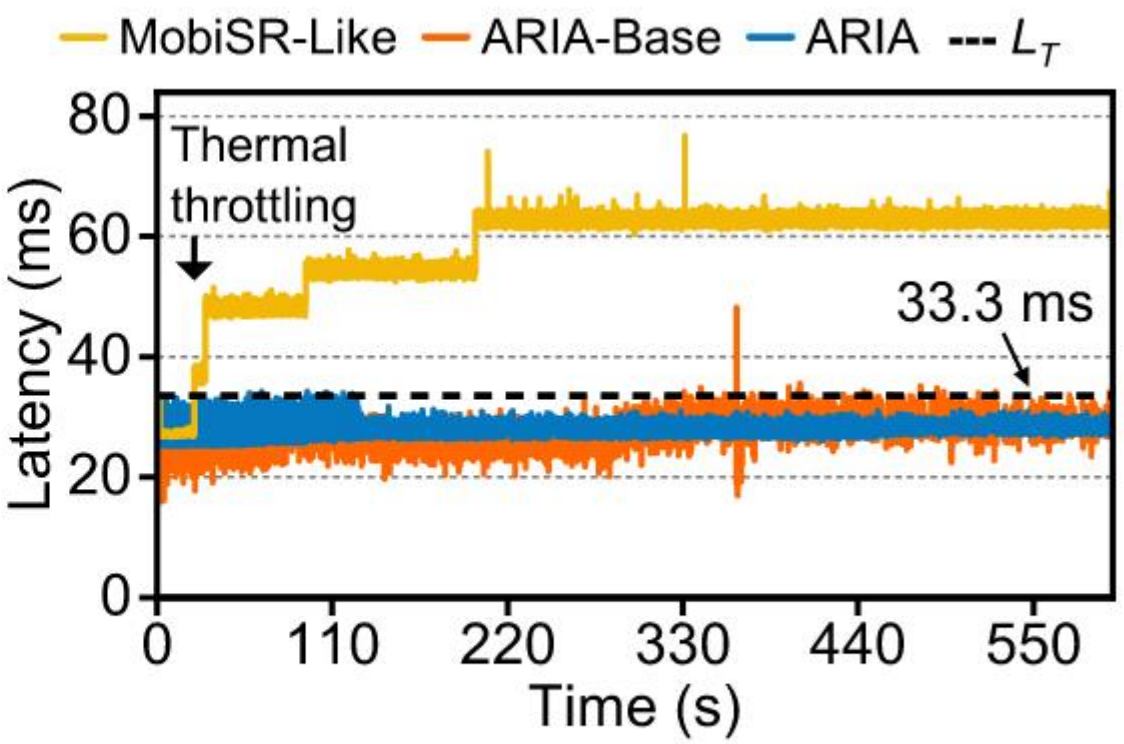
- **硬件平台:** 三星 Galaxy S22/S24 手机 (搭载高通 8Gen1/8Gen3芯片)
- **数据集:** 自建的大规模 MARV 数据集，包含 8.3 万张带有多样化场景、相机运动、深度和语义真值的图像
- **Baseline:**
  NPU-Only: 只用 NPU 低分辨率跑
  MobiSR-Like: 一种已有的将图像分块给不同处理器的方法
  ARIA-Base: ARIA 的一个简化基线版本，它实现了 GPU+NPU 协同工作的基本思想，但缺少所有核心的优化模块 。它使用简单的图像块匹配来识别动态区域 ， 并用 2D 平面变换和双线性插值来做结果的对齐与融合

实验评估

核心结果 I：延迟与 Deadline 达成率 (DSR)



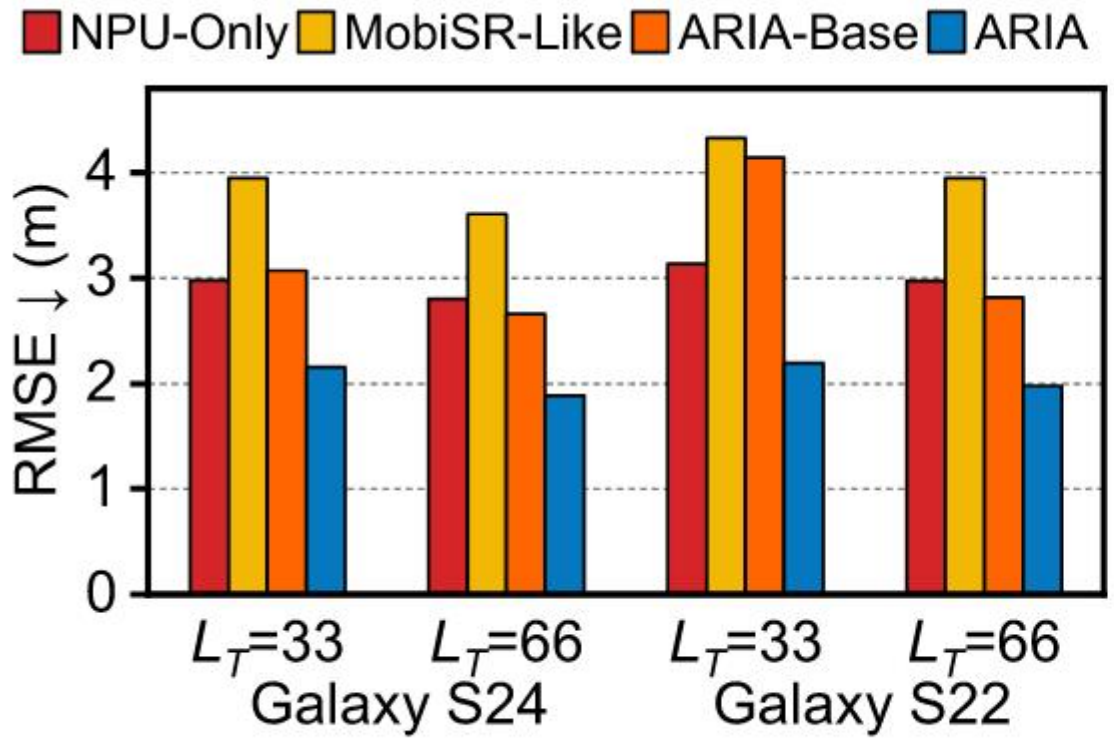
(a) Deadline success rate



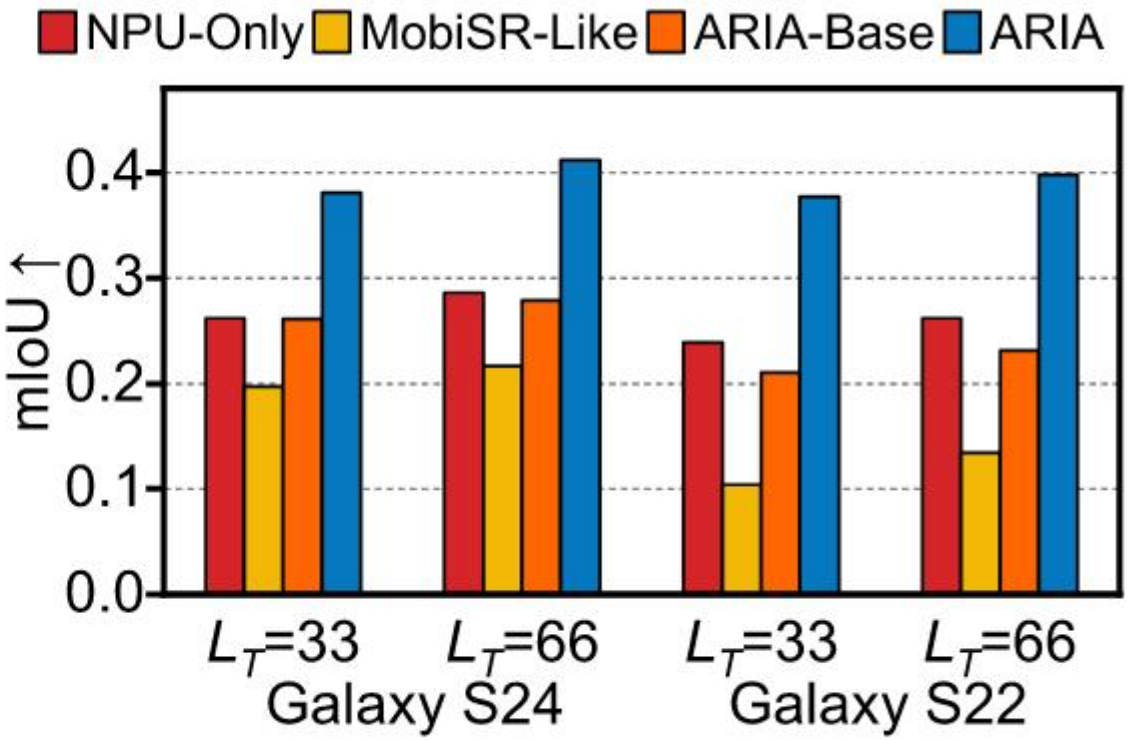
(b) End-to-end latency on S24

实验评估

核心结果 II：预测精度



(a) Metric depth estimation



(b) Semantic segmentation

实验评估

核心结果 II：预测精度

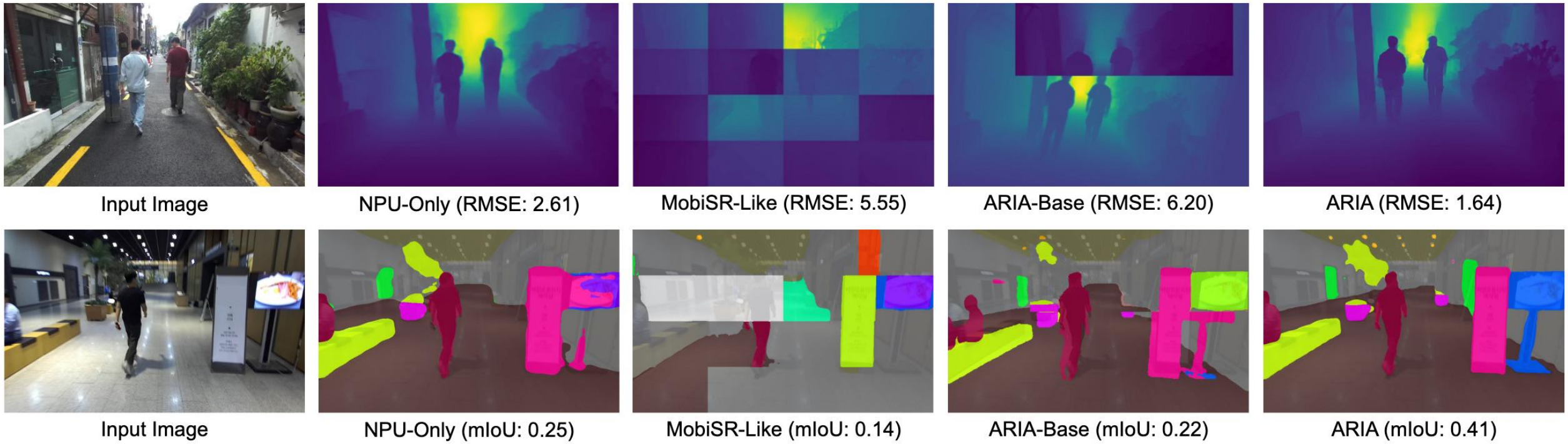


Figure 14: Example VFM predictions on MARV dataset with Samsung Galaxy S24 under $L_T = 33.3$ ms.

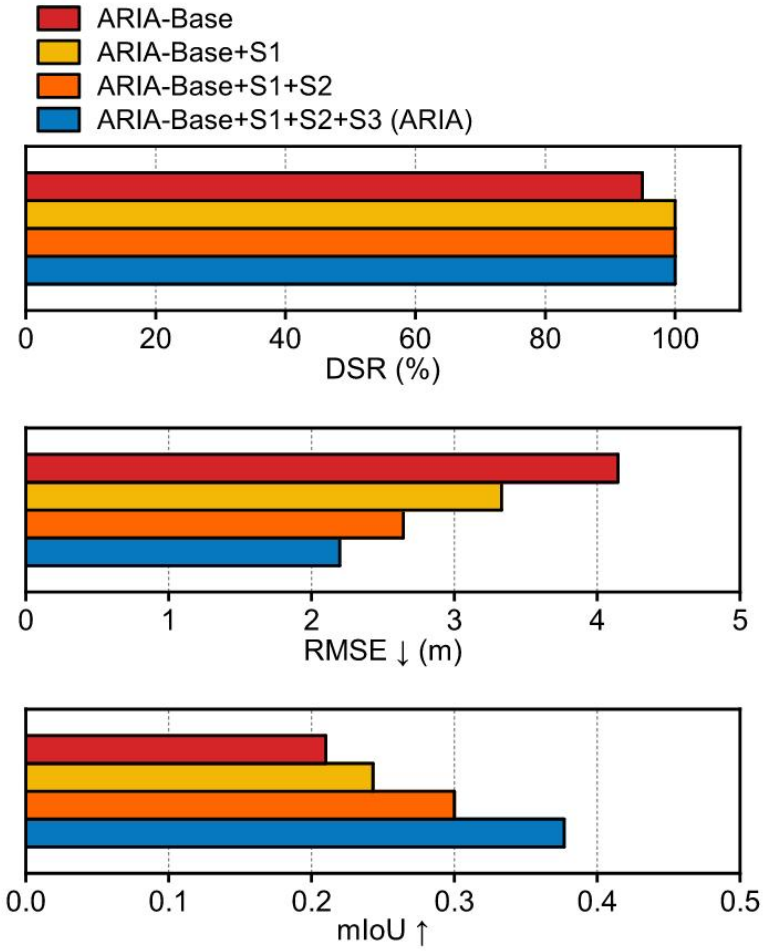
实验评估

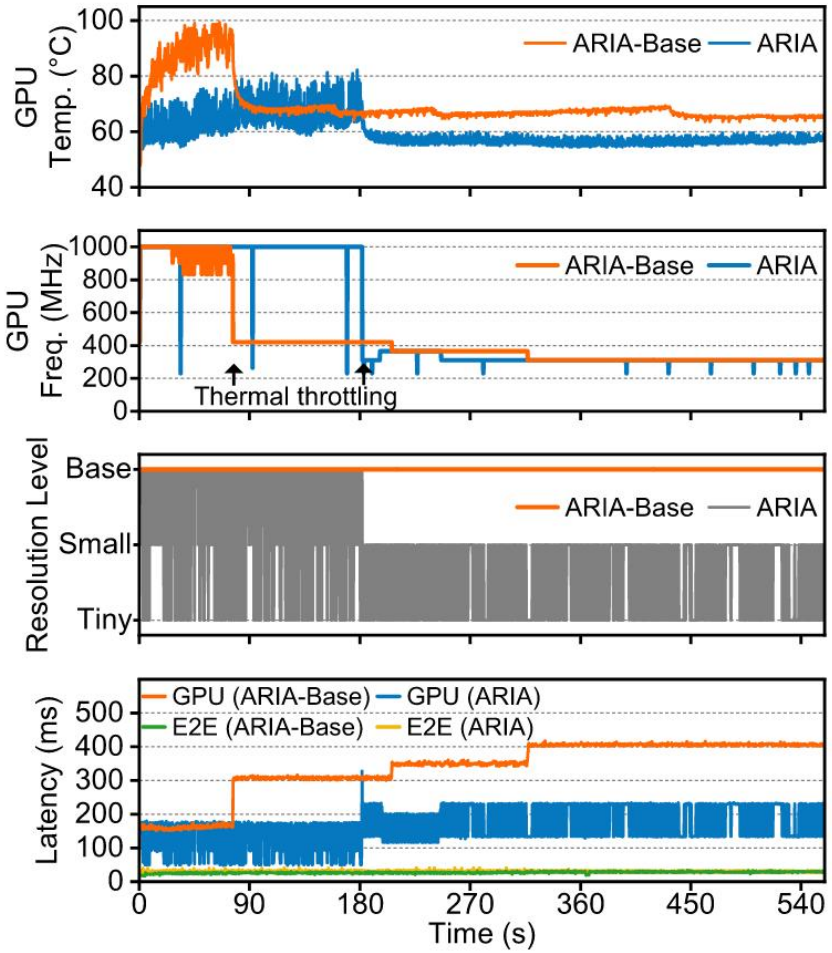
消融实验

S1: 动态区域识别与跟踪

S2: 时空对齐模块

S3: 执行器调度模块





章节介绍

- 作者介绍
- 研究背景与动机
- ARIA 系统设计
- 实验评估
- **总结与讨论**

总结与讨论

总结:
ARIA 通过一种并行与选择性的推理框架，结合了 GPU 的高质量全局处理能力和 NPU 的高效率局部更新能力，提高了 VFM 在移动设备上的推理速度，满足了移动 AR 应用对实时性和精度上的要求

能否提高:
ARIA 的实验和方案设计，很大程度上依赖于移动 NPU 具备高效的浮点（FP16）运算能力，以避免 VFM 模型在整型（INT8）量化下严重的精度损失 。如果目标设备的 NPU 只擅长 INT8 运算，ARIA 方案的有效性可能会打折扣，这在一定程度上限制了其在更广泛异构硬件上的普适性

我们的工作: 针对大语言模型端侧复杂应用的场景优化，将高负载的 Prefill/Rerank 任务切分到 GPU 和 NPU上并行计算