湖 北 省 软 件 企 业 协 会
团 体 标 准

T/HBSEA 016—2025

AI 端侧设备算力模型适配测试规范

Adaptation testing specifications for computing power and models of AI end devices

2025-02-27 发布　　2025-02-28 实施

湖 北 省 软 件 企 业 协 会　发 布

目　录

前　言

本标准按照GB/T1.1-2020《标准化工作导则第1部分：标准化文件的结构和起草规则》的规定起草。

本标准由湖北省软件企业协会提出并归口。

本标准起草单位：武汉国创超算科技有限公司、华中科技大学、天翼云科技有限公司湖北分公司、上海兆芯集成电路股份有限公司、武汉金美创新科技有限公司、上海云源创信息技术有限公司、深圳市鸿普森科技股份有限公司、上海智众联电子销售有限公司、缘边智联(南京)智能科技有限公司、湖北公众信息产业有限责任公司、天翼电信终端有限公司湖北分公司、湖北省公安科学技术研究所、湖北省高级人民法院、湖北省软件企业协会。

本标准主要起草人：赵德亚、胡晓娅、李建坤、齐子时、文筠、陈晶、骆吉波、左三化、黄河、张发胜、胡可、王崇鲁、熊伟、刘晓、田野、骆斌、李琪、温晖、朱丽、赵秋迪、高鹤鸣、徐杨。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本标准于2025年2月首次发布。

引　言

为全面客观衡量 AI 端侧设备算力与模型的适配性，构建一套科学的、系统的、有效的评估模型和方法，特制定本标准。

本标准为企业开展 AI 端侧设备算力模型适配测试提供了测试指标体系和测试方法，可用于发现企业在推进 AI 端侧设备算力模型适配工作中存在的优势和不足，明确进一步改进的方向，为 AI 端侧设备部署相关人工智能应用提供决策参考。同时，也可以用于第二方或第三方在测试、采购选型中对 AI 端侧设备算力模型适配性进行评估。

AI端侧设备算力模型适配测试规范

1 范围

本文件规定了AI端侧设备的算力与其可承载的AI模型的适配性测试方法。

本文件适用于AI端侧设备算力模型适配性评估。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，标注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271. 1-2000 信息技术词汇第 1 部分:基本术语

GB/T 11457-2006 信息技术 软件工程术语

YD/T 3944-2021 人工智能芯片基准测试评估方法

T/CESA 1121-2020 人工智能芯片 面向端侧的深度学习芯片测试指标与测试方法

T/CESA 1169-2021信息技术 人工智能 服务器系统性能测试规范

3 术语、定义和缩略语

3.1 术语和定义

3.1.1

端侧场景 scene of terminal side

以数据获取及本地处理为主要特征的场景。

3.1.2

端侧设备 device of terminal side

布置在端侧场景以数据获取及本地处理为主要功能的设备，如摄像头、移动通信终端、机器人、无人机、可穿戴设备等。

3.1.3

人工智能 Artificial Intelligence

表现出人类智能(如推理和学习)相关的各种功能的功能单元和能力。

[来源:GB/T 5271.28-2001，28.01.02]

3.1.4

深度学习 Deep Learning

机器学习中一种对数据进行表征学习的方法。通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。

[来源:ISO/ECTR 29119-11:2020(en)，3.1.26]

3.1.5

基准测试 Benchmark

通过设计科学的测试方法、测试工具和测试系统，实现对一类测试对象的某项性能指标进行定量的和可对比的测试。

3.1.6

推理 inference

在机器学习中，推理通常指将训练过的模型应用于无标签样本，进而来做出预测的过程。

3.1.7

批次 batch

模型训练的一次迭代(即一次梯度更新)中使用的样本集。

3.1.8

批次大小 batch size

一个批次中的样本数。批次大小在训练和推理期间通常是固定的。

3.1.9

样本 sample

数据集中用于训练或推理运算的单个对象，可以是一副图片或一句语句。

3.1.10

作业 workload

一组被一同送入训练或推理系统的 N 个样本，N 为正整数。

注:例如，单次作业包含了 8 张图片。

3.2 缩略语

下列缩略语适用于本文件：

AI	人工智能	Artificial Intelligence
BLEU	双语评估替补	Bilingual Evaluation Understudy
DUT	被测设备	Device Under Test
FLOPs	浮点运算数	floating point of operations
FPS	每秒处理帧数	Frame Per Second
MAC	乘累加单元	Multiply and Accumulate
mAP	均值平均精度	Mean Average Precision
NMS	非极大值抑制	Non-maximum Suppression
OPs	操作数	OPerations
ROC	受试者工作特征曲线	Receiver operating characteristic curve

4 测试项目概述

测试项组成及依赖关系总体要求如下：

——测试项分为 2 大类共 5 个子项，包含 AI 端侧设备的基础能力测试、性能测试。

——2 大类测试项是逐层递进关系，如果前面的测试项没有通过，后面的测试项无法继续进行。

测试项列表如表 1 所示，具体测试步骤及预期结果详见本标准第 6 章。

表 1 测试项列表

序号	测试项分类	测试项	测试成果
1	基础能力测试	1.1 设备基础测试	处理器、RAM、存储容量、电池容量等
		1.2 场景测试	具体使用场景等
		1.3 基本功能测试	通信速率、读写速度、带宽、吞吐量等
2	性能测试	2.1 训练测试	训练时长、功耗、吞吐率、资源利用率等
		2.2 推理测试	推理时长、推理功耗、吞吐率等

5 测试通用要求

5.1 测试框架

AI 端侧设备算力模型适配测试框架如图 1 所示：

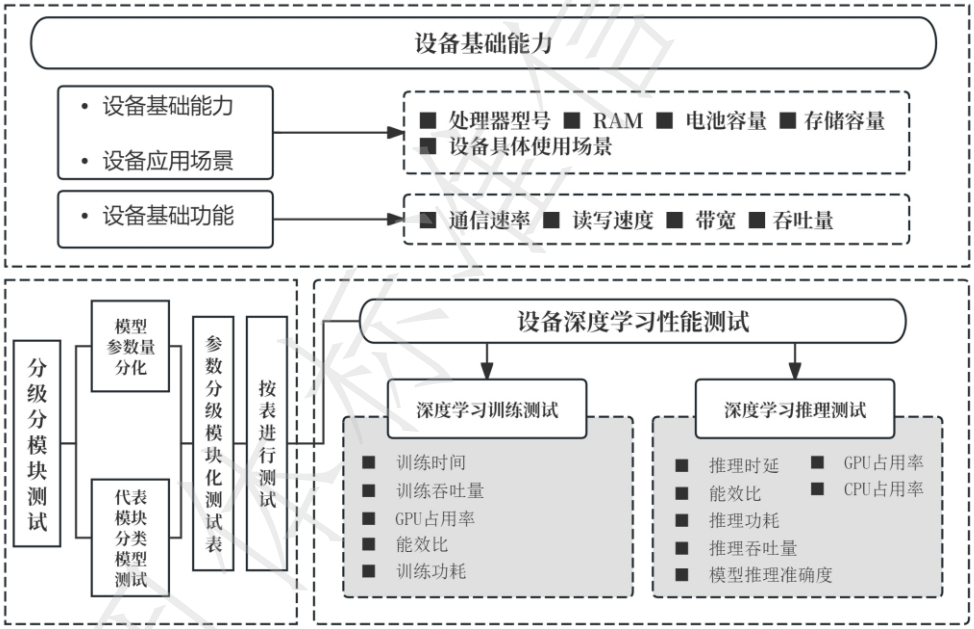

图 1 算力模型适配测试框架图

5.2 测试原则

5.2.1 实用性

测试方法的设计应能够产生积极效果，对实际应用具有指导意义。

5.2.2 公平性

基准测试方法必须依据明确的规则和指标，以确保不同对象之间的比较是公正的。

5.2.3 客观性

在进行测试时，应始终坚持以客观的科学数据为依据，确保评测的公正性。

5.2.4 可重复性

测试应保证在不同测试环境下对同一对象进行测试时，能够得到一致的测量结果。

5.3 AI端侧设备基础参数

AI 端侧设备基本信息评估阶段主要通过材料审查的方式来验证参评设备基本信息的真实性和完整性。在此过程中，将重点关注设备的基本信息，包括设备名称、基本描述、功能说明等关键参数。

评估准则如下：

必选项目：企业必须提交规定的材料以供审查，这些材料通常包括但不限于设备的技术规格书、功能说明书以及其他证明设备性能和特性的官方文件。

可选项目：企业可以根据自身情况选择性提交相关材料进行审查，这些材料可能包括额外的性能测试报告、用户手册、案例研究等，旨在提供更全面的设备性能和应用场景说明。

需提供材料如表 2 所示。

表 2 AI 端侧设备测试材料检查表

项目	是否必选	提交材料
AI 端侧设备基本信息		
AI 端侧设备名称、版本号	必选	信息介绍
AI 端侧设备功能说明	必选	同上
AI 端侧设备外形及尺寸	必选	同上
AI 端侧设备功耗情况	必选	同上
支持的操作系统及版本	必选	同上
支持的深度学习框架	必选	同上
知识产权状况说明	可选	同上
行业实施案例	可选	介绍相关应用情况

5.4 测试场景

在设计适用于 AI 端侧设备算力模型适配的测试用例时，必须考虑到不同参数量和计算量对处理器在计算、存储和通信方面的影响。因此，选择具有代表性

的网络模型进行测试是至关重要的。下述的评估场景和网络模型作为测试的参考。

鉴于技术不断进步和被测设备之间的差异，测试场景的选择应灵活，以适应具体的测试环境和需求。这意味着测试方案需要根据实际情况进行定制化调整，以确保测试的准确性和有效性。这种灵活性允许测试团队针对不同的技术迭代和设备特性，优化测试流程，从而更准确地评估神经网络处理器的性能。

5.4.1 图像分类

任务描述:识别图像中的物体类别。

代表模型：MobileNet vl，MobileNet v2，ResNet 50 等。

数据集：ImageNet。

参考评价标准:计算指定精度下的分类任务的 Accuracy。

5.4.2 目标检测

任务描述：在给定的图像中精确定位物体并标注其类别。

代表模型：Faster R-CNN、Yolo v11、MobileNet+SSD、Mask R-CNN、SSD。

数据集：VOC2012 或 COCO2017。

参考评价标准：通过计算 mAP、IoU、NMS 来评估目标检测任务的精度。

5.4.3 超分辨率

任务描述：恢复给定缩小版（如 4 倍缩小）的原始照片。

代表模型：VDSR。

数据集：VOC2012。

参考评价标准：使用 PSNR 和 SSIM 作为性能评价指标。

5.4.4 图像语义分割

任务描述：根据图像中语义含义的不同，对像素进行分组或分割。

代表模型：Deeplabv3+。

数据集：VOC2012 或 Cityscapes。

参考评价标准：使用 IoU 作为性能评价指标。

5.4.5 机器翻译

任务描述：将一种自然语言转换为另一种自然语言。

代表模型：Seq2Seq、BERT、Transformer 等。

数据集：Wikipedia 或 WMT English-German。

参考评价标准：使用 BLEU 作为翻译任务的评估指标。

5.4.6 自然语言处理

任务描述：对人类语言的理解和生成，进行情感分析、文本摘要、问答系统等。

代表模型：BERT、GPT、Transformer-XL 等。

数据集：GLUE、SQuAD、Wikipedia 等。

参考评价标准：准确率（Accuracy）、精确度（Precision）、召回率（Recall）、F1 分数（F1 Score）等。

5.4.7 多模态

任务描述：处理和理解来自多种模态（如文本、图像、声音等）的信息，以实现跨模态的交互和理解。

代表模型：CLIP、ViLBERT、LXMERT 等。

数据集：MSCOCO、Flickr30k、Conceptual Captions 等。

参考评价标准：准确率（Accuracy）、平均精度均值（mAP）、BLEU 分数等。

5.5 测试环境部署

测试环境部署如图 2 所示：

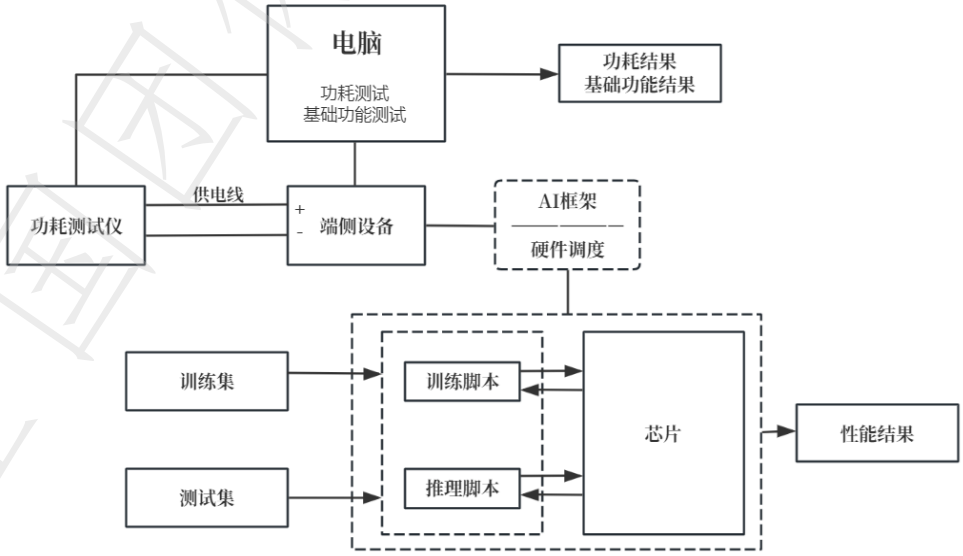

图 2 测试环境示意图

将电脑通过接口与端侧设备相连，同时示波器的正负极也与端侧设备相连接。

利用基准测试工具，对端侧设备时间性能和评估指标进行了测试。此外，功耗分析软件工具被用来分析示波器反馈的电流和电压结果，从而对端侧设备的功耗进行评估。

AI 框架是支持人工智能技术实现的软件架构，它们包括但不限于深度学习平台，如 TensorFlow 和 PyTorch，以及专为移动设备或推理任务设计的框架，例如 TensorFlow Lite 和 TensorRT。这些框架为构建和部署深度学习模型提供了必要的工具和库。需结合具体端侧设备部署支持框架。硬件调度具备双重功能：一方面，它能够兼容上层的人工智能框架；另一方面，它能够管理和分配包括中央处理器（CPU）、图形处理器（GPU）、数字信号处理器（DSP）以及神经网络处理器（NPU）在内的硬件资源，以满足人工智能计算的需求。

5.6 分级分模型测试说明

鉴于各类 AI 端侧设备在芯片架构和性能上存在显著差异，这些差异直接影响了设备能够支持的神经网络模型的参数规模。不同架构的神经网络模型在性能表现上也呈现出显著的异质性。分级分模型测试用于系统性地评估和对比不同参数量级以及不同架构的神经网络模型的性能表现。这种方法的核心在于两个维度：参数量分级和架构模块分模型。

(1) 参数量级划分：根据模型的参数规模，即模型中可训练参数的总数，将模型划分为不同的量级。这一指标是衡量模型复杂度和学习能力的关键因素，对于确定模型在特定硬件上的可行性和效率至关重要。

(2) 架构与模块分析：首先按照模型的整体架构进行分类，例如区分卷积神经网络（CNN）、Transformer 等主流架构。继而在每种架构内部，根据模型的主要模块或组件进行进一步的细分，以深入探究不同架构组件对模型性能的具体影响。

具体分级可参照附表 3，4 所示。

6 测试步骤

6.1 基础能力测试

6.1.1 设备基础参数测试

测试编号	**1-1**
测试项目	AI 端侧设备基础参数检验

测试编号	**1-1**
测试目的	获取设备相关参数，对设备进行基础分级
测试方式	材料审查
前置条件	相关硬件参数可查，已提供设备说明书
测试步骤	1)阅读相关说明书获取，获取相关硬件参数信息； 2)在系统中进行查看、调用，检查是否与说明书描述一致
预期结果	正确获取 AI 端侧设备的处理器规格、RAM、存储容量、电池容量等参数

6.1.2 设备应用场景测试

测试编号	**1-2**
测试项目	AI 端侧设备场景应用测试检验
测试目的	验证 AI 端侧设备在特定应用场景下的性能和功能是否满足预定的规范和标准
测试方式	人工审查
前置条件	已提供 AI 端侧设备相关应用场景
测试步骤	1)确认 AI 端侧设备说明书、技术文件、应用案例； 2)与测试场景实际数据进行对比，评估设备是否适用该场景
预期结果	AI 端侧设备能够基本满足该测试场景中的应用

6.1.3 AI端侧设备基础功能测试

测试编号	**1-3**
测试项目	AI 端侧设备基本功能测试
测试目的	测试 AI 端侧设备的存储能力、网络能力等基础能力
测试方式	人工审查
前置条件	AI 端侧设备已通电并处于正常工作状态，测试工具和环境已准备就绪

测试步骤	1) 测试 AI 端侧设备的通信速率、读写速度、带宽、吞吐量等参数；
预期结果	获取 AI 端侧设备的基本功能的相关参数数据

6.2 算力测试

6.2.1 训练测试

测试编号	**2-1**
测试项目	AI 端侧设备模型训练测试
测试目的	验证 AI 端侧设备在执行 AI 模型训练任务时的性能
测试方式	遍历测试
前置条件	AI 端侧设备已通电并处于正常工作状态，必要的训练数据已加载
测试步骤	1) 被测者按测试内容，编写并运行必要的训练代码（包含数据预处理、数据读入、训练、结果模型格式转化与持久化），得到结果 2) 训练期间，记录过程数据、按规定测量、计算指标值、记录日志、生成结果数据； 3) 测试者检查结果合规性；
预期结果	AI 端侧设备在 AI 模型训练任务中表现出预期的性能

6.2.2 推理测试

测试编号	**2-2**
测试项目	AI 端侧设备模型推理测试
测试目的	验证 AI 端侧设备在执行 AI 模型推理任务时的性能
测试方式	遍历测试
前置条件	AI 端侧设备已通电并处于正常工作状态，必要的推理数据已加载
测试步骤	1) 测试 AI 端侧设备在不同批次大小下的推理性能； 2) 记录过程数据，按规定测量、计算指标值、记录日志、生成结果数据； 3) 测试者检验结果合规性
预期结果	AI 端侧设备在 AI 模型推理任务中表现出预期的性能

附录

参数量分级见分级表 3。

表 3 模型参数量分级表

等级	参数量范围（M）	特点	适用场景
第一级	0 - 2	小规模模型模型结构简单，计算量和存储需求最低。适合嵌入式设备和实时应用。	移动设备、智能摄像头、实时监控系统
第二级	2 - 10	中小规模模型在资源消耗和性能之间取得较好平衡。适用于大多数边侧 AI 应用。	移动端应用、智能家居设备、便携式设备
第三级	10 - 50	中等规模模型提供更强的特征提取能力和更高的准确率。适用于对性能有较高要求但资源仍受限的场景。	高级图像识别、智能监控、复杂的边缘计算任务
第四级	50 - 100	中大型模型具备更高的模型复杂度和表达能力。适用于高端边侧设备或需要处理复杂任务的应用。	高分辨率图像处理、复杂的语义分割、高性能边缘服务器
第五级	100-500	大型模型提供最强的性能和特征表达能力，但对计算资源和存储空间要求极高。	高级视觉任务、大规模数据处理、高精度需求的工业应用
第六级	500-1000	超大型模型具有极高的特征提取和推理能力，适用于需要高精度和复杂推理的场景。	自然语言处理中的大规模模型、超高分辨率图像生成、复杂科学计算
第七级	1000-5000	支持更高维度的表达和复杂推理，对硬件要求极高，适用于超高性能计算场景。	超大规模语言模型（如 GPT-4+）、高精度预测建模
第八级	5000+	计算量和存储需求超越传统硬件能力，需专用硬件支持，适用于未来的智能和推理前沿领域。	下一代人工智能系统、跨领域大规模协作任务、超高精度科学模拟

模型分类见分类表 4。

表 4 模型分类表

主要网络架构	代表模块	功能介绍	代表模型
CNN	普通卷积（Vanilla Convolution）	基本的二维卷积操作，用于提取图像的局部特征。	LeNet5, AlexNet, VGG
	密集连接（Dense Connectivity）	每一层与前面所有层相连，增强特征复用和梯度传播。	DenseNet
	残差连接（Residual Block）	通过跳跃连接缓解深层网络中的梯度消失问题，促进训练。	ResNet, Wide-ResNet
	组卷积（Group Convolution）	将输入通道分组进行卷积，减少计算量并提升并行度。	ResNeXt
	Squeeze-and-Excitation (SE) 模块	通过全局信息重新校准通道权重，增强特征表达能力。	SEResNet, SEResNeXt, MobileNetV3
	倒转瓶颈+深度可分离卷积（Inverted Bottleneck + Depthwise Separable Convolution）	通过倒转瓶颈结构和深度可分离卷积实现高效特征提取。	MobileNetV2, EfficientNet, EfficientNetV2, ConvNeXt
	通道混洗（Channel Shuffle）	在组卷积后打乱通道顺序，促进跨组信息交流。	ShuffleNetV1, ShuffleNetV2
	重参数化卷积（Re-parameterizable Convolution）	训练时使用复杂结构，推理时转换为简化的卷积结构，提高推理效率。	RepVGG, RepLKNet
	多尺度处理（Multi-scale Processing）	同时处理多种尺度的特征，增强模型的多尺度感知能力。	Res2Net
	高分辨率融合（High-resolution Fusion）	在多个分辨率下保持高分辨率特征图	HRNet
	交叉阶段部分网络（Cross Stage Partial, CSPNet）	通过部分层的跨阶段连接减少计算量，同时保持特征表达能力。	CSPNet
	正则化卷积块（Regularized Convolution Blocks）	采用正则化策略设计卷积块，提升模型的泛化能力和可扩展性。	RegNet

Transformer	多头自注意力（Multi-Head Self-Attention）	通过多个注意力头并行计算，实现对不同子空间特征的捕捉。	Vision Transformer (ViT), DeiT, BEiT
	窗口自注意力（Window-based Self-Attention）	将注意力计算限制在局部窗口内，提升计算效率和扩展性。	Swin-Transformer, Swin Transformer V2, Twins
	嵌套 Transformer（Nested Transformers）	通过嵌套结构增强模型的表达能力。	Transformer-in-Transformer
	基于 MLP 的混合（MLP-based Mixing）	使用多层感知机进行特征混合，替代部分自注意力机制。	MLP-Mixer
	Meta-Former 结构	基于 Meta-Former 框架，统一不同类型的算子模块设计。	PoolFormer
	层次化注意力（Hierarchical Attention）	通过层次化设计提升对多尺度特征捕捉能力。	HorNet, VAN
	高效 Transformer 设计（Efficient Transformer Designs）	采用优化的注意力机制和结构设计，提升计算效率和模型性能。	EfficientFormer, EVA, MViT V2, MobileViT, DaViT