基于改进 AdaTAD 的奶山羊爬跨行为时序动作定位方法

王佳源[1], 李其同[1], 罗元滔[1], 杨蜀秦[2], 王振华[1], 宁纪锋[1*], 王美丽[1]

(1. 西北农林科技大学信息工程学院,陕西杨凌 712100,中国; 2. 西北农林科技大学机械与电子工程学院,陕西杨凌 712100,中国)

摘 要: [**目的/意义**] 奶山羊爬跨行为的时序定位是繁殖管理的重要基础。针对现有方法多停留在行为判别层面、在未修剪视频中对短时突发行为的起止边界刻画不足,且易受遮挡、视角变化与背景干扰影响的问题,提出一种基于面向时序动作定位的适配器调优(Adapter Tuning for Temporal Action Detection, AdaTAD)改进的端到端时序动作定位方法,以实现爬跨行为的准确识别与起止时间精确定位。[**方法**] 以 AdaTAD 框架为基线,引入视觉提示调优,通过少量可学习 Prompt Tokens 对主干注意力分布进行任务引导,增强关键帧及边界邻域的特征响应;设计多尺度运动感知适配器,采用并联多尺度时序深度可分离卷积分支建模不同时间尺度的运动模式,并结合残差连接与非线性映射稳定注入主干特征,提升短时微动作与相对完整动作过程的联合建模能力。[**结果和讨论**] 所提方法的平均精度均值达到 81.72%,相较基准模型 AdaTAD 提升 5.00 个百分点;在时间交并比为 0.7 的更严格条件下达到 68.85%,较 AdaTAD 提升 4.06 个百分点,表明该方法在高边界精度要求下仍具有优势。模型推理速度为 65.78 帧/s,可训练参数量为 27.941 M,在精度提升的同时保持较低开销。[**结论**] 该方法可提升复杂养殖场景下奶山羊爬跨行为的时序定位精度与稳定性,为繁殖行为监测与管理决策提供关键时序信息支撑。

关键词: 爬跨行为;时序动作定位;AdaTAD;精准养殖;多尺度运动建模;端到端模型;智慧养殖

中图分类号: S24;TP391.4 **文献标识码:** A **文章编号:** SA202601012

引用格式:王佳源, 李其同, 罗元滔, 杨蜀秦, 王振华, 宁纪锋, 王美丽. 基于改进 AdaTAD 的奶山羊爬跨行为时序动作定位方法[J]. 智慧农业(中英文), 2026, 8(4): 148-163. DOI: 10.12133/j.smartag.SA202601012

WANG Jiayuan, LI Qitong, LUO Yuantao, YANG Shuqin, WANG Zhenhua, NING Jifeng, WANG Meili. Temporal Action Localization of Mounting Behavior in Dairy Goats Based on Improved AdaTAD[J]. Smart Agriculture, 2026, 8(4): 148-163. DOI: 10.12133/j.smartag.SA202601012 (in Chinese with English abstract)

0 引 言

近年来,随着羊乳及其制品需求的不断增长,奶山羊养殖业在全球范围内受到越来越多的关注[1]。与此同时,养殖模式正经历由传统的小规模分散饲养向规模化、集约化养殖转型的过程,对高效繁殖管理提出了更高要求。智能养羊技术的发展为提升繁殖效率和生产力提供了新的解决思路[2]。奶山羊爬跨行为是其繁殖活动中的关键环节,对其进行准确、及时的监测不仅能够提高配种效率,还可有效降低空怀率,从而显著提升整体养殖效益[3]。在此背景下,通过自动化手段对奶山羊爬跨行为进行精准识别和定位,可为养殖场制定科学的繁殖管理策略提供重要依据,从而有效提高受孕率与产羔率。

进一步从智慧养殖管理的角度看,奶山羊爬跨行为不仅是发情状态的重要外在表征,也是配种管理与繁殖决策中的关键信息来源。在实际生产流程中,若能够在未修剪视频中准确定位爬跨行为的发生时间及持续区间,并与羊只身份识别结果相结合,可形成“个体级、时间级”的繁殖行为记录,如每只羊是否发生爬跨、爬跨时间及爬跨频次等关

收稿日期:2026-01-09

基金项目:国家重点研发计划项目(2022YFD1300200);陕西秦创原引用高层次创新创业人才项目(QCYRCXM-2022-359)

作者简介:王佳源,硕士研究生,研究方向为基于计算机视觉的奶山羊爬跨行为时序动作定位方法研究。E-mail:yolo_wjy@nwafu.edu.cn

*通信作者:宁纪锋,博士,教授,研究方向为计算机视觉及机器学习。E-mail:njf@nwsuaf.edu.cn

copyright©2026 by the authors (登录 www.smartag.net.cn 免费获取电子版全文)

键指标。这有助于及时识别未发生爬跨或爬跨异常的个体，为补配或健康检查提供决策参考，降低漏配与空怀风险，提升繁殖管理的精准性。

在上述繁殖管理需求背景下，传统的爬跨行为监测主要依赖人工观察，然而这种方式存在人力成本高、实时性差以及易受主观因素影响等局限。可穿戴传感器虽然能提供精细的行为数据，但会干扰动物自然行为，并增加设备成本[4]。相比之下，基于计算机视觉等非接触式的畜牧发情和爬跨行为检测方法以其非侵入性、自动化和高效性逐渐成为研究热点[5]。WANG等[6]提出利用双通道声学标签结合长短期记忆网络判别模型与小波阈值自适应滤波，实现复杂环境下母牛发情声音的高精度检测。WANG等[7]开发了奶牛发情识别YOLO模型（Estrus- You Only Look Once, E-YOLO），通过引入归一化瓦瑟斯坦距离（Normalized Wasserstein Distance, NWD）损失函数，显著提升了视觉模型对目标位置偏差的敏感性。WANG等[8]则基于改进YOLOv5s设计轻量化奶牛爬跨识别系统，通过网络剪枝与注意力机制的结合，实现了检测精度与推理效率的平衡。随后，SHI等[9]采用时序增强三维卷积网络（Temporal-Boost 3D Convolutional Network）结合尾部关键点检测与动态加权融合，在群养条件下实现了母羊发情期摆尾行为的精准识别。DUAN等[10]通过改进声纹特征提取与分类网络结构，有效提升了母猪发情声音在噪声环境下的识别性能。CAO等[11]进一步融合多类声学特征，并结合数据增强与Transformer时序建模，在母猪发情音频及多个公开数据集上展现出优异的鲁棒性和泛化能力。与此同时，WANG等[12]将改进YOLOv8n与目标跟踪模型结合，实现了奶牛爬跨行为的高精度检测。此外，从智慧畜牧研究的整体研究视角看，相关综述对基于视觉与智能感知的牲畜行为监测、健康评估及繁殖管理研究进行了系统总结，指出在复杂养殖环境下，实现稳定、连续的行为时序建模仍面临一定挑战[13]。

尽管上述研究验证了深度学习方法在牲畜发情和爬跨行为监测中的有效性，但现有方法大多侧重于判断行为是否发生，对于爬跨行为的具体时间边界缺乏系统研究，限制了其在精细化繁殖管理中的应用潜力。奶山羊爬跨行为具有高度随机性和短时性，这对模型的时序建模能力和特征提取效率提出了更高要求。时序动作定位作为视频理解领域的重要研究方向，其目标是从未修剪的视频中同时识别动作类别并精确定位动作的起止时间[14]。近年来已涌现出基于上下文建模与图结构关系建模等多种代表性方法[15-17]。此外，随着端到端检测框架的发展以及时序动作检测与视频时刻检索统一建模研究的推进，该任务受到了广泛关注[18]。将时序动作定位方法引入奶山羊爬跨行为监测，不仅能够实现精细化时序分析，也为提升养殖管理水平和经济效益提供了新的技术支持。

本研究提出一种端到端的奶山羊爬跨行为时序动作定位方法，该方法以面向时序动作定位的适配器调优方法[19]（Adapter Tuning for Temporal Action Detection, AdaTAD）为基础模型，通过视觉提示调优（Visual Prompt Tuning, VPT）[20]机制引入少量可学习的提示词元（Prompt Tokens），对主干网络特征提取进行任务驱动式轻量引导，从而增强模型对关键帧的聚焦能力。同时设计了多尺度运动感知适配器（Multi-Scale Motion Adapter, MSMA），通过并联多尺度深度可分离卷积自适应捕捉不同节奏与幅度的运动模式，结合残差连接与非线性激活函数，提升短时微动作和长时间动作的联合建模能力。实验结果表明，该方法能够在视频序列中精确识别奶山羊爬跨行为并定位动作的起止时间，为真实养殖场景下繁殖行为的自动感知与量化分析提供可靠的时序信息基础，从而为后续发情监测、繁殖管理及异常行为分析等智慧养殖决策环节提供有效的数据支撑。

1　材料与方法

1.1　数据集构建与分析

1.1.1　数据采集与基本特征

本研究所用数据采集自中国陕西省咸阳市杨陵区西北农林科技大学畜牧学教学与实验基地，实验对象为饲养于该基地的2～5岁萨能奶山羊。数据采集时间为2024年6—7月，该时段处于奶山羊的主要配种期，母羊发情表现较为集中、爬跨行为频繁，为获取具有代表性的爬跨行为样本提供了良好的自然条件。为保证视频质量，采用索尼FDR-AX100E型摄像机和索尼A5000型摄像机拍摄，共获取142段未修剪的奶山羊爬跨行为视频序列。

本研究涉及的奶山羊养殖环境主要由饮食区与运动区构成，其典型环境示例如图1所示。其中，饮食区以室内圈舍为主，环境相对封闭，栏杆、饲喂设施等结构性遮挡较为常见，如图1a所示；运动区位于室外，为奶山羊日常活动与自由行走区域，

背景开阔但个体间交互频繁，易出现羊只相互遮挡等情况，如图1b所示。饮食区与运动区空间上相邻，奶山羊可通过多扇小门在两类区域之间自由出入，使得同一视频序列中可能同时包含室内与室外场景动态转换。



a. 饮食区环境 b. 运动区环境

图1 奶山羊生活环境

Fig. 1 Living environment of dairy goats

数据采集采用手持摄像机拍摄。受养殖场通行空间、栏舍结构与动物活动位置变化等因素影响，摄像机安装高度与拍摄角度未做统一固定，而是在保证目标个体清晰可见、关键行为过程可完整覆盖的前提下，根据现场情况对拍摄位置进行调整，以保证不同活动区域内行为片段的有效记录。因此，数据在可见性约束下包含一定范围的视角与尺度变化，并伴随不同区域背景条件带来的复杂性差异。在上述拍摄方式与养殖环境条件下，数据集中包含了不同拍摄视角、成像尺度及复杂遮挡条件下的爬跨行为样本。其典型行为样本示例如图2所示，涵盖室外远景与室内圈舍环境，并包括无遮挡、羊只遮挡及设施遮挡等多种复杂场景条件。



a. 室外远景的侧面爬跨行为 b. 室外群羊遮挡的爬跨行为

c. 室内栏杆遮挡的侧面爬跨行为 d. 室内栏杆遮挡的正面爬跨行为

图2 奶山羊爬跨行为数据集示例

Fig. 2 Examples of the dataset for mounting behavior in dairy goats

从行为时序特性来看，奶山羊爬跨行为持续时间整体较短，多数集中在 1～10 s，而对应的未修剪视频长度分布相对分散，主要集中在 2～40 s，最长可达 187 s。单个未修剪视频中包含的爬跨行为实例数量不等，经人工标注统计，以“单段视频中独立爬跨行为实例的个数”为统计单位，平均为 1.4 次/视频，在个别持续时间较长、群体活动较为频繁的视频中，爬跨行为可多次发生，单段视频中最多可标注出 10 次彼此时间上相互独立的爬跨行为实例。上述特征体现了数据在时间长度和行为频次上的多样性，为评估模型在不同背景比例和行为密度条件下的时序定位能力提供了数据基础。

1.1.2　行为标注流程与规范

为保证奶山羊爬跨行为时序标注的准确性与一致性，本研究在数据集构建过程中制定了统一的人工标注规范，并采用牛津大学视觉几何组开发的视觉标注工具（Visual Geometry Group Image Annotator, VIA）对未修剪的视频进行时序标注。VIA 是一种轻量级、开源的视频与图像标注工具，支持在时间轴上对行为片段进行精确标记，能够同时记录行为类别及其对应的起止时间。

在正式标注前，标注人员（具备养殖行为基础认知）对奶山羊爬跨行为的判定标准与时间边界定义进行了统一说明。标注过程中，针对每段未修剪视频，标注人员首先在 VIA 的时间轴界面中对爬跨行为片段进行定位，并为每个行为实例标注其行为类别（爬跨），以及对应的起始时间与结束时间，从而形成完整的行为时序标注结果，如图 3 所示。

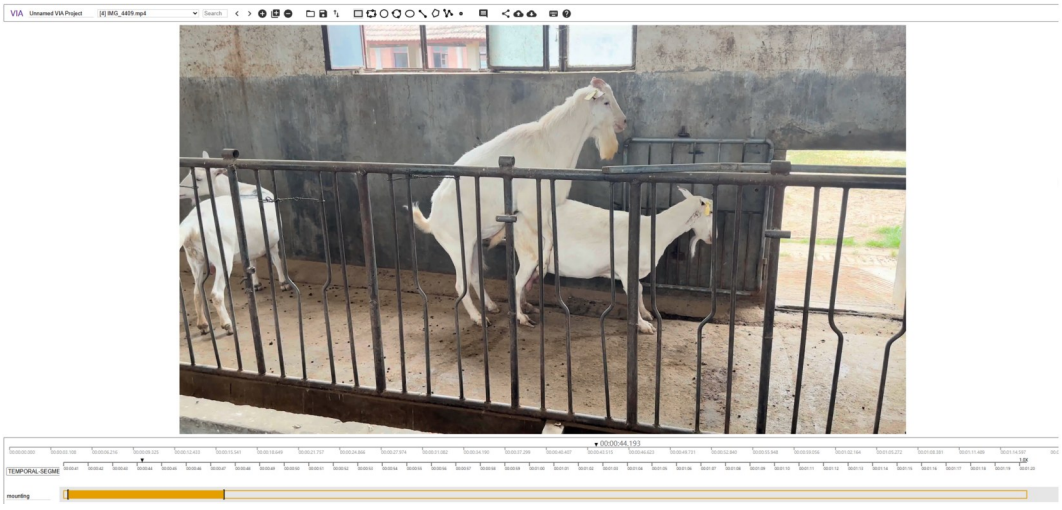


图 3　基于 VIA 工具的奶山羊爬跨行为时序标注示例

Fig. 3　Examples of temporal annotation of mounting behavior using VIA

在本研究中，爬跨行为的起始时间定义为公羊出现明确爬跨意图并开始实施前肢抬起、跃起或跨越等动作的首帧；结束时间定义为爬跨动作完成后，公羊前肢完全落地并与被爬跨羊体态明显分离的末帧。对于起止时间边界不完全清晰或动作存在过度模糊的情况，以视频中可观察到的关键动作变化作为主要判定依据，并尽量保持标注时间边界的一致性。

在具体标注过程中，每段视频均由至少一名标注人员逐帧查看并完成标注；当存在时间边界判断分歧或行为判定不一致的情况时，由多名标注人员共同复核并协商确定最终标注结果，以尽量降低人工主观判断带来的偏差。通过上述标注流程与规范设置，尽可能保证奶山羊爬跨行为起止时间标注的一致性与可靠性。

1.1.3　数据集划分与分布统计

考虑到奶山羊爬跨行为具有持续时间短、发生随机性强等特点，且单段未修剪视频中目标行为与背景片段比例差异较大，本研究以视频序列为基本单位进行数据集划分，从而避免同一视频内的不同行为实例同时出现在训练与测试阶段所导致的信息泄露，保证评估的独立性与公平性。

在真实养殖场景的小样本研究背景下，若进一步从有限视频中单独划分验证集，将导致训练样本规模进一步缩小，不利于模型对不同视角、遮挡与背景复杂度条件下爬跨行为多样性的学习。因此，本研究采用训练集/测试集的二分划分方式对 142 段未修剪视频进行划分：其中 65 段视频用于模型训练；77 段视频作为固定测试集仅用于最终性能评估。在模型结构与训练超参数设定过程中，相关设

置参考原始AdaTAD工作及先验经验确定；固定测试集在整个过程中不参与任何形式的参数选择与模型调优，以降低未单独设定验证集可能引入的评估偏倚风险。同时，为降低单次划分带来的偶然性影响，本研究在训练集内部补充开展5折稳定性评估实验以验证性能稳定性（见2.6节）。

为呈现二分划分后训练集与测试集在关键场景因素上的构成情况，本研究在遮挡类型（无遮挡、羊只遮挡、设施遮挡）与拍摄视角（正面、侧面、后方）两个维度对两部分数据进行了统计，结果见表1。可以看出，两部分数据在各组合下均有样本覆盖，未出现某一组合完全缺失的情况。该统计用于描述并核查二分划分后的数据分布概况，为后续实验评估提供分布层面的依据。

表1 训练集与测试集在遮挡类型与拍摄视角下的视频数量

Table 1 Video-level distribution of the training and test sets across occlusion types and camera views

单位:个

遮挡情况	数据集	正面	侧面	后方	合计
无遮挡	训练集/测试集	5/6	6/10	4/4	15/20
羊只遮挡	训练集/测试集	6/6	9/10	5/6	20/22
设施遮挡	训练集/测试集	5/6	16/18	9/11	30/35
合计	65/77	16/18	31/38	18/21	142

需要说明的是，数据采集于真实养殖场环境，部分奶山羊个体可能在不同时间段的视频中重复出现。本研究的目的是评估模型在真实养殖环境中对爬跨行为的时序动作定位能力，因为该设置与实际场景一致。

此外，基于人工标注得到的行为起止时间，本研究以行为实例为统计单位对数据集中全部爬跨行为实例的持续时间进行了统计分析。结果表明，单段爬跨行为具有显著的短时性特征：平均持续时间为4.74 s，中位数为3.47 s，四分位区间（P25～P75）为2.12～5.95 s，且86.39%的行为实例持续时间分布在［1，10］ s，同时存在少量长尾实例（最大23.67 s）。上述统计结果客观刻画了奶山羊爬跨行为“持续时间短、起止边界变化快且时长分布存在一定离散性”的时序特征：短时行为对边界回归精度提出更高要求，而时长离散性与长尾现象也提示模型需要具备多尺度时序建模能力以适应不同节奏与幅度的动作过程，从而为本研究采用端到端时序动作定位框架及多尺度运动建模设计提供数据层面的支撑。

1.2 研究方法

1.2.1 AdaTAD时序动作定位方法

时序动作定位（Temporal Action Detection, TAD）任务旨在从未修剪视频中同时识别目标行为类别并精确定位其起始与结束时间，对模型的时序建模能力和边界判别能力提出了较高要求。AdaTAD是一种端到端的时序动作定位框架，其在冻结大规模主干网络参数的前提下，引入参数高效微调机制和轻量级时序信息适配器（Temporal-Informative Adapter, TIA），以降低训练开销并增强时序建模能力[19]。

其中，TIA通过在时间维度嵌入深度可分离卷积，提升模型对动作时序结构的感知能力。鉴于AdaTAD在建模规模、迁移效率及参数效率等方面的综合优势，本研究选取其作为基础时序动作定位模型，并针对奶山羊爬跨行为持续时间短、动作幅度小且具有突发性的特点进行改进，以提升模型在复杂养殖场景下的时序定位精度与稳定性。

AdaTAD的整体流程包括视频主干特征提取、时序检测头预测和后处理3个环节：主干网络对原始视频帧进行时空特征编码，随后采用无锚框（Anchor-free）的时序检测头对各时间位置进行分类与边界回归，并通过后处理得到最终的动作时间段预测。后文统一以“AdaTAD框架”指代该端到端流程。

1.2.2 提出的方法

尽管AdaTAD在通用时序动作定位任务中表现良好，但将其直接应用于奶山羊爬跨行为定位仍面临一定挑战。奶山羊爬跨行为通常持续时间较短，且关键动作帧（如起始跃起、前肢抬起和结束落地）在视觉上与非目标行为差异有限，容易受到背景运动、个体遮挡及视角变化的干扰。

从特征建模过程来看，原始AdaTAD架构中缺乏对主干网络注意力分布的显式引导机制，在训练初期模型容易在长时间序列中对特征进行均匀建模，导致对爬跨行为关键起止帧的响应不够集中，从而影响时序边界定位的准确性与稳定性。

此外，AdaTAD所采用的TIA模块主要基于单一时间尺度进行特征建模，难以同时刻画爬跨行为中短时微动作与相对完整动作过程之间的多尺度动态差异，限制了模型对复杂时序模式的表达能力。

基于上述分析，本研究从“关键帧显式引导”和“多尺度运动建模”两个互补角度对原始Ada-

TAD 模型进行改进：

1）引入 VPT 机制，通过轻量化的提示向量引导主干网络更加关注与爬跨行为密切相关的关键时序区域；

2）设计 MSMA，以增强模型对不同时间尺度动作模式的联合建模能力。

其中，VPT 侧重于在全局时序范围内强化模型对潜在爬跨关键帧的关注，而 MSMA 则在此基础上进一步刻画不同时间尺度下的动作演化过程，二者在功能上形成互补。该设计使模型在不显著增加参数量和计算复杂度的前提下，同时提升了爬跨行为的候选片段召回能力与时间边界回归精度，更符合实际养殖场景中对离线或准实时行为监测在稳定性与效率方面的综合需求。改进后的整体网络结构如图 4 所示。

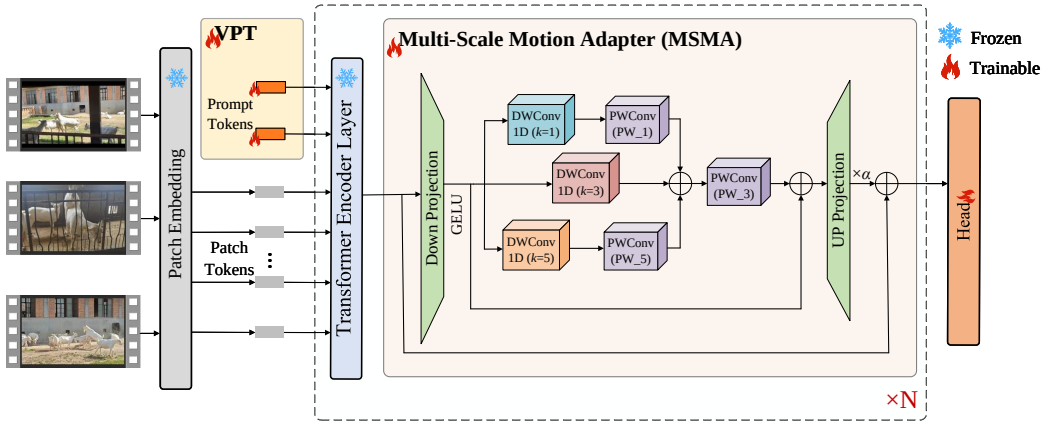


图 4　改进 AdaTAD 奶山羊爬跨行为时序动作定位模型结构图

Fig. 4　Architecture of the improved AdaTAD model for temporal action localization of mounting behavior in dairy goats

在上述结构基础上，本研究模型沿用 AdaTAD 的检测头与后处理流程，在推理阶段输出动作实例集合，如公式（1）所示。

$$\mathcal{Y}=\{(t_s^k,\ t_e^k,\ c^k,\ s^k)\}_{k=1}^{K} \tag{1}$$

式中：$\mathcal{Y}$表示模型输出的动作实例集合；K为该未修剪视频中预测得到的动作实例数量；t_s^k与t_e^k分别表示第k个动作片段在原始视频时间轴上的起始与结束位置，可进一步换算为秒或帧索引；c^k为动作类别，本研究为“爬跨”单类别；s^k为对应的置信度评分。上述输出由 AdaTAD 框架内部的 Anchor-free 时序检测头产生：检测头在时间轴上的每个候选位置同时预测类别置信度与边界回归量，并解码得到候选时间段；随后对候选时间段进行置信度阈值筛选与时间域非极大值抑制（Temporal NMS），以去除冗余的重叠预测，最终得到动作时间段集合$\mathcal{Y}$。需要说明的是，本研究的改进仅作用于主干特征编码阶段（引入 VPT 与 MSMA）以提升特征表征质量，检测头结构及后处理流程保持与原始 AdaTAD 框架一致。

1.2.3　Visual Prompt Tuning 机制

在奶山羊爬跨行为时序动作定位任务中，模型不仅需要识别爬跨行为的存在性，更需要对其起始与结束时间进行精确定位。然而，由于该行为持续时间短、关键动作帧占比低，且在视觉上与非目标行为差异有限，传统基于主干网络特征提取的建模方式难以在长时间序列中稳定聚焦于爬跨行为的关键时序区域。

为此，本研究引入 VPT 机制，以轻量化的方式在 Transformer 中插入可学习的提示向量（Prompt Tokens），用于引导主干网络更加聚焦于与目标行为相关的关键时序区域。这些提示向量作为任务相关的条件化引导信号，在多层 Transformer 编码器中与所有 Patch Tokens 建立注意力交互，从而逐层聚合全局语义并动态调整特征提取的侧重点。通过这种方式，Prompt Tokens 能够在主干网络冻结的情况下，有效增强模型对爬跨行为关键帧的感知能力，提升时序动作定位的准确性与鲁棒性。尤其是在时序动作定位任务中，该机制有助于在长时间序列中放大与爬跨行为起止位置相关的判别性特征响应，从而为后续时间边界回归提供更加稳定的特征基础。

为简洁起见，本研究在公式推导中省略批次维度（Batch Dimension），仅给出单样本张量形状；在实际实现中，公式（2）对应张量形态为$\boldsymbol{X}_i \in \mathbb{R}^{b\times(p+m)\times d}$，其中$b$为批次大小。

在第i层 Transformer 编码器输入处，将p个可

学习 Prompt Tokens 与当前层输入的 Patch Tokens 进行拼接，构成输入，形式如公式（2）所示。

$$\boldsymbol{X}_i = \left[\boldsymbol{P}_i;\ \boldsymbol{E}_{i-1}\right], \boldsymbol{X}_i \in \mathbb{R}^{(p+m)\times d} \tag{2}$$

式中：$\boldsymbol{X}_i$ 表示第 i 层 Transformer 编码器的输入特征；$\boldsymbol{P}_i \in \mathbb{R}^{p\times d}$ 表示第 i 层输入处引入的独立可学习 Prompt Tokens；$\boldsymbol{E}_{i-1} \in \mathbb{R}^{m\times d}$ 表示当前层输入的 Patch Tokens 表达；p 为 Prompt 个数；d 为特征维度。

需要说明的是，m 表示单次送入主干网络的定长视频片段在 Patch/Tubelet Embedding 后产生的 Patch Tokens 数量，其取值由输入分辨率及空间 Patch 大小与时间 Tubelet 大小共同决定，为固定常数，不随原始未修剪视频总时长变化。不同长度视频在进入主干网络前已通过定长窗口（训练阶段随机截取/补齐、测试阶段滑动窗口推理并融合）统一为相同形状的输入，因此单次前向的 token 数保持一致。

拼接后的特征序列输入第 i 层 Transformer，经过多头自注意力机制更新后如公式（3）所示。

$$[\boldsymbol{Z}_i, \boldsymbol{E}_i] = L_i([\boldsymbol{P}_i;\ \boldsymbol{E}_{i-1}]) \tag{3}$$

式中：L_i 表示第 i 层 Transformer 编码器的映射操作；$\boldsymbol{Z}_i \in \mathbb{R}^{p\times d}$ 和 $\boldsymbol{E}_i \in \mathbb{R}^{m\times d}$ 分别为本层输出的 Prompt Tokens 特征表达和 Patch Tokens 特征表达。需要特别说明的是：依据 VPT 的设计，每 1 层使用的 Prompt Tokens 参数 $\boldsymbol{P}_i$ 是相互独立的可学习参数，并在该层输入前插入。虽然第 i 层会产生更新后的 prompt 表达 $\boldsymbol{Z}_i$，但在进入第 $i+1$ 层时，$\boldsymbol{Z}_i$ 不作为下一层输入继续传递，而是被新的独立参数 $\boldsymbol{P}_{i+1}$ 替换；只有 Patch Tokens 表达 $\boldsymbol{E}_i$ 在层间持续传递并累积更新。该设计使不同层的 prompts 能够分别学习该层所需的任务引导信号。

在注意力机制中，Prompt Tokens 能够与所有 Patch Tokens 建立全连接关系，其注意力计算表达如公式（4）所示。

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\times \boldsymbol{V} \tag{4}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示查询矩阵、键矩阵和值矩阵；Softmax（·）表示归一化函数。

通过多层编码过程中的注意力交互，Prompt Tokens 能够在特征编码阶段对注意力分配产生任务条件化引导，从而提升模型对关键动作帧及边界附近时序变化的响应强度，增强后续候选评分与时间边界回归所依赖的判别性特征质量。在本研究设置中，在主干前 6 层 Transformer 中插入 Prompt Tokens，数量设置为 p=2，使模型在保持主干冻结的同时仍能获得关键行为的显式引导与更强的行为感知和时序建模能力。

从时序动作定位流程的角度来看，VPT 作用于主干特征编码阶段，并将其影响传递至后续的候选片段评分与时间边界回归模块。通过在全局时间范围内显式引导模型关注与爬跨行为相关的关键帧区域，VPT 有助于提升潜在爬跨片段的可分性与召回能力，同时减弱大量背景帧对边界判别的干扰，从而为精确起止时间定位提供更加可靠的特征基础。

1.2.4 Multi-Scale Motion Adapter

在奶山羊爬跨行为时序动作定位任务中，目标动作往往具有突发性、持续时间短且幅度有限的特点，边界过渡不明显；同时受拍摄视角、距离与遮挡影响，动作演化节奏在不同场景下存在显著差异。若仅采用单一时间尺度建模，模型容易在“短时微动作”与“相对完整动作过程”的表征之间产生偏差，从而降低候选片段评分与起止边界回归的稳定性。为此，本研究提出 MSMA，以轻量化方式在时间维度捕捉多尺度动态变化，并通过残差路径稳定注入主干特征，从而增强复杂场景下的边界判别能力。

1）输入张量与符号约定。为便于与 Transformer 输出对齐，MSMA 的输入为编码器输出的 Token 序列，表示为 $\boldsymbol{x} \in \mathbb{R}^{B\times N\times d}$，其中 B 为批次大小；N 为 Token 数；d 为特征维度。因此，Transformer Encoder Layer 的输出可表示为 N 个 Token，每个 Token 的维度为 d。在本模块中，进一步将 Token 序列视作来自定长视频片段的时空 Patch/Tubelet 表示，并令公式（5）：

$$N = t \times h \times w \tag{5}$$

式中：t 为时间维度的 Token 数（即 Tubelet 在时间轴展开后的长度）；h，w 为空间维度的 Token 网格尺寸（由输入分辨率与 Patch 大小决定，均为常数）。在实际实现中，t，h，w 由主干 Patch/Tubelet Embedding 的输出形状确定，并在 MSMA 前向过程中以参数形式传入用于张量重排。

2）通道压缩与非线性映射。MSMA 首先通过下投影全连接层对通道维度进行压缩，并引入高斯误差线性单元激活函数（Gaussian Error Linear Unit, GELU）以增强表达，如公式（6）所示。

$$\boldsymbol{z} = \sigma(\boldsymbol{x}\boldsymbol{W}_{\text{down}}), \boldsymbol{W}_{\text{down}} \in \mathbb{R}^{d\times d_\gamma}, d_\gamma = \lfloor r\cdot d \rfloor \tag{6}$$

式中：$\sigma(\cdot)$ 为 GELU 激活函数；$r \in (0,\ 1)$ 为

压缩比例；$\boldsymbol{z} \in \mathbb{R}^{B \times N \times d_\gamma}$表示下投影后的特征张量；$\boldsymbol{W}_{\text{down}}$表示下投影权重矩阵；$d_\gamma$为压缩后的通道维度。因此，下投影层对序列中每个Token共享参数进行线性变换与激活，其输出形状与输入Token数一致，仅通道维度由d变为d_γ。

3）时空重排与时序深度可分离卷积。为在时间维度建模动作演化，MSMA将$\boldsymbol{z}$从Token序列重排回时空结构$\left[B|t|h|w|d_\gamma\right]$，并对每个空间位置$(h|w)$的时间序列执行1D卷积。具体记重排后的张量为$\tilde{\boldsymbol{z}}$，并进一步展开空间维度得到公式（7）。

$$\tilde{\boldsymbol{z}} \in \mathbb{R}^{(Bhw) \times d_\gamma \times t} \tag{7}$$

式中：(Bhw)表示对每个空间位置形成一条长度为t的时序特征序列。

随后，MSMA在时间轴上构建3个并行的时序（Temporal）深度可分离1D卷积分支，卷积核大小$k \in \{1, 3, 5\}$用于捕捉不同时间尺度的动态变化。记$\text{DWConv}_k(\cdot)$为沿时间维度执行的逐通道一维卷积（Groups=d_γ），$\text{PW}_k(\cdot)$为后接的逐点卷积（Point-Wise 1 × 1 Convolution）用于通道混合，如公式（8）所示。

$$\begin{cases} \boldsymbol{y}_1 = \text{PW}_1(\text{DWConv}_1(\tilde{\boldsymbol{z}})) \\ \boldsymbol{y}_5 = \text{PW}_5(\text{DWConv}_5(\tilde{\boldsymbol{z}})) \end{cases} \tag{8}$$

式中：$\boldsymbol{y}_1$和$\boldsymbol{y}_5$分别表示卷积核大小为1和5的时序卷积分支输出。为保证不同卷积核的输出在时间维度上长度一致，本研究将卷积步长设为1，并使用同长度填充（same padding）策略：对奇数核k取padding = dilation $\cdot (k-1)/2$。因此三分支输出均保持相同的时间长度t，可进行逐元素融合。

4）跨尺度残差注入与瓶颈内残差。不同于直接对三分支输出进行简单求和，MSMA将短期与长期分支的结果以残差形式注入到中期分支（$k=3$）以完成跨尺度聚合，如公式（9）所示。

$$\Delta\tilde{\boldsymbol{z}} = \text{PW}_3(\text{DWConv}_3(\tilde{\boldsymbol{z}}) + \boldsymbol{y}_1 + \boldsymbol{y}_5) \tag{9}$$

并将$\Delta\tilde{\boldsymbol{z}}$逆重排回token形状$\Delta\boldsymbol{z} \in \mathbb{R}^{B \times N \times d_\gamma}$，与瓶颈特征进行残差叠加，如公式（10）所示。

$$\boldsymbol{z}' = \boldsymbol{z} + \Delta\boldsymbol{z} \tag{10}$$

式中：$\Delta\tilde{\boldsymbol{z}}$表示时空重排空间中的跨尺度融合特征；$\Delta\boldsymbol{z}$表示逆重排后的token序列增量特征；$\boldsymbol{z}'$表示瓶颈内残差更新后的特征表示。该“瓶颈内残差”用于稳定梯度传播，并保留原始压缩特征的信息主干。

5）上投影与外部残差注入。最后，MSMA通过上投影层恢复原始通道维度，并以带可学习缩放因子α的残差形式注入输入特征，如公式（11）所示。

$$\boldsymbol{x}' = \boldsymbol{x} + \alpha \cdot (\boldsymbol{z}'\boldsymbol{W}_{\text{up}}),\ \boldsymbol{W}_{\text{up}} \in \mathbb{R}^{d_\gamma \times d}, \tag{11}$$

式中：$\boldsymbol{x}'$表示经MSMA调整后的输出特征；$\boldsymbol{W}_{\text{up}}$表示上投影权重矩阵；$\alpha$为可学习缩放因子，用于自适应控制MSMA分支对主干特征的注入强度。由此，MSMA在参数开销较小的情况下实现了多尺度时序动态建模，并通过残差路径提升动作边界附近特征响应的判别性与稳定性。

1.3 实验平台和参数设置

实验所用显卡是NVIDIA TESLA A40，操作系统为Ubuntu 20.04，深度学习框架为PyTorch 2.0.1，并基于CUDA 11.8与cuDNN 8.2.4加速计算。训练过程中采用AdamW优化器，迭代60个epoch，批处理量为2，初始学习率设为1×10^{-4}，并结合线性预热与余弦退火策略进行动态调整。对于非端到端模型，实验使用在Kinetics-710数据集上预训练的VideoMAEv2-g提取离线特征；而本研究提出的端到端模型的主干网络采用VideoMAEv2-s，初始化权重使用在Kinetics-400数据集上预训练得到的公开模型参数进行训练与微调。为保证实验对比的公平性，各模型均在相同的数据划分和训练配置下进行评估。

1.4 评价指标

本研究采用时间交并比（Temporal Intersection over Union, tIoU）与平均精度均值（mean Average Precision, mAP）@［0.3：0.1：0.7］对TAD性能进行评估。需要指出的是，TAD中的预测结果与真实标注均表示为一维时间段（Temporal Segment），即动作在视频时间轴上的起止时间$[t_{\text{start}}, t_{\text{end}}]$。

1）tIoU。用于衡量预测动作片段与真实动作片段在时间维度上的重合程度。对同一视频内预测段$\hat{y} = [\hat{t}^s, \hat{t}^e]$与真实段$y = [t^s, t^e]$，其定义如公式（12）所示。

$$\text{tIoU}(\hat{y}, y) = \frac{\max\left(0, \min(\hat{t}^e, t^e) - \max(\hat{t}^s, t^s)\right)}{(\hat{t}^e - \hat{t}^s) + (t^e - t^s) - \max\left(0, \min(\hat{t}^e, t^e) - \max(\hat{t}^s, t^s)\right)} \tag{12}$$

式中：$\hat{y}$表示预测动作时间段，y表示真实动作时间段；$\hat{t}^s$和$\hat{t}^e$分别表示预测动作片段的起始时间和结束时间；t^s和t^e分别表示真实动作片段的起始时间和结束时间。tIoU取值范围为[0，1]，值越大

表示预测时间段与真实时间段重合越充分、边界定位越精确。不同tIoU阈值可视为对起止时间误差容忍度的不同要求，其中较高阈值（如tIoU ≥ 0.7）对动作边界回归精度提出更严格约束。

2）平均精度（Average Precision, AP）的计算。在每个阈值$\tau \in T = \{0.3, 0.4, 0.5, 0.6, 0.7\}$下，首先将预测片段按置信度从高到低排序。对排序后的每个预测片段，仅与同一视频中的真实片段计算tIoU，并按一对一匹配约束判定真阳性（True Positive, TP）/假阳性（False Positive, FP）：若存在尚未被匹配的真实片段使得tIoU ≥ τ，则该预测记为TP，并锁定对应真实片段（保证在同一阈值下每个真实片段最多匹配1次）；否则记为FP。随后根据累计TP/FP构建Precision-Recall曲线，并沿用AdaTAD评测脚本计算该阈值下的平均精度$\mathrm{AP}(\tau)$。

3）mAP@tIoU。本研究为单类别时序动作定位任务，因此对任意阈值τ如公式（13）所示。

$$\mathrm{mAP}(\tau) \equiv \mathrm{AP}(\tau) \tag{13}$$

综合指标为跨阈值平均，如公式（14）所示。

$$\text{avg mAP} = \frac{1}{|T|}\sum_{\tau \in T}\mathrm{mAP}(\tau) = \frac{1}{|T|}\sum_{\tau \in T}\mathrm{AP}(\tau) \tag{14}$$

式中：$T = \{0.3, 0.4, 0.5, 0.6, 0.7\}$表示评价时采用的tIoU阈值集合，$|T|$表示集合$T$中阈值的个数。该指标能够同时反映模型在不同时间重合要求下的检测精度与边界回归能力，避免单一阈值评价带来的偶然性偏差，因此被广泛用于TAD任务的综合性能评估。

2 结果与分析

2.1 不同模型的比较实验

为了验证所提出模型在奶山羊爬跨行为时序动作定位任务中的有效性，本研究以跨1 000帧且具有10亿参数的端到端时序动作检测方法AdaTAD[19]为基线模型，并选取具有代表性的时序动作定位方法作为对比模型，在奶山羊爬跨行为时序动作定位数据集上进行实验，包括基于Transformer的端到端时序动作检测[21]（Temporal Action Detection with Transformer, TadTR）、用于时序动作定位的视频自拼接图网络[22]（Video Self-Stitching Graph Network for Temporal Action Localization, VSGN）、用于无锚框时序动作定位的显著边界特征学习方法[23]（Learning Salient Boundary Feature for Anchor-free Temporal Action Localization, AFSD）、基于Transformer的动作时刻定位方法[24]（Localizing Moments of Actions with Transformers, ActionFormer）、基于相对边界建模的时序动作检测方法[25]（Temporal Action Detection with Relative Boundary Modeling, TriDet）、用于时序动作检测的动态特征聚合方法[26]（Dynamic Feature Aggregation for Temporal Action Detection, DyFADet）、通过重构预训练视频骨干网络实现可逆时序动作定位方法[27]（Rewiring Pretrained Video Backbones for Reversible Temporal Action Localization, Re2TAL），实验结果如表2所示。

表2 不同模型的奶山羊爬跨行为时序动作定位实验结果

Table 2 Experimental results of temporal action localization for mounting behavior in dairy goats with different models

单位：%

模型类别	模型	avg mAP	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7
非端到端模型	TadTR	42.90	56.09	52.65	42.42	36.05	27.32
	VSGN	47.89	68.28	58.97	49.18	40.32	22.68
	ActionFormer	77.63	86.42	85.00	79.60	73.39	63.74
	TriDet	78.89	88.25	87.81	79.33	74.32	64.74
	DyFADet	80.52	91.36	90.11	79.89	76.99	64.23
端到端模型	AFSD	56.43	81.76	73.13	60.93	45.61	20.69
	AdaTAD	76.72	87.10	80.90	78.49	72.32	64.79
	Re2TAL	75.66	86.71	82.96	77.28	71.17	60.17
	改进AdaTAD	81.72	90.46	86.70	83.47	79.11	68.85

注：由于包含非端到端方法，参数量/测试时间强依赖离线特征提取器设置，故本研究统一在表3报告端到端链路下（改进AdaTAD与AdaTAD）的参数量与帧率作为参考。

表2按是否采用端到端（End-to-End, E2E）训练/推理范式将方法划分为两组。本研究对E2E的划分依据为训练与推理链路是否以原始视频帧为输入并进行联合优化，而非网络内部结构层面的阶段划分；两类方法的差异主要体现在输入表征形式（原始帧vs.离线特征）与优化方式（联合学习vs.固定特征上的学习）上。为减少离线特征差异对非端到端方法结果的影响，表2中非端到端方法统一采用在Kinetics-710上预训练的VideoMAEv2-g提取离线特征；本研究方法属于端到端范式，主干采用VideoMAEv2-s，并使用在Kinetics-400上预训练的公开权重初始化。跨范式效率指标的统计口径差异已在表注中说明，并在表3进一步给出端到端链路下（本研究方法与AdaTAD）的参数量与帧率（Frames Per Second, FPS）作为参考。

从实验结果可以看出，AdaTAD在对比模型中取得了较为稳定的性能表现，验证了其作为基准模

表3 基于AdaTAD改进的奶山羊爬跨行为时序动作定位模型消融实验结果

Table 3 Ablation results of the improved AdaTAD-based temporal action localization model for dairy goat mounting behavior

模型	FPS/（帧/s）	可训练参数量/M	总参数量/M	avg mAP/%	tIoU=0.3/%	tIoU=0.4/%	tIoU=0.5/%	tIoU=0.6/%	tIoU=0.7/%
AdaTAD	66.19	27.703	49.583	76.72	87.10	80.90	78.49	72.32	64.79
AdaTAD + MSMA	58.51	27.936	49.815	81.00	88.64	86.59	82.56	77.73	69.47
AdaTAD + VPT	66.80	27.708	49.587	79.64	91.56	85.21	79.73	75.18	66.53
改进AdaTAD	65.78	27.941	49.820	81.72	90.46	86.70	83.47	79.11	68.85

型应用于本研究的合理性。在此基础上，本研究提出的改进模型在各个tIoU阈值下均取得了更优的检测性能。

具体而言，本研究方法在tIoU为0.5、0.6和0.7时的mAP值以及平均mAP值均优于其他8种时序动作定位模型。其平均mAP为81.72%，较基线AdaTAD提升5.00个百分点，表明基于VPT机制和多尺度运动感知适配器的改进AdaTAD模型能够有效提升奶山羊爬跨行为时序动作定位的准确度。

此外，在更严格的高阈值条件下，本方法同样表现突出。在tIoU=0.7时，mAP达到68.85%，较AdaTAD提升4.06个百分点，进一步说明所提模型在动作边界回归精度和定位准确性方面具有优势。

综上，本研究提出的模型不仅摆脱了对离线特征的依赖，而且在整体精度和高阈值条件下均优于现有代表性方法，验证了其在奶山羊爬跨行为时序动作定位任务中的有效性和优势。

2.2 模型整体消融实验

表3给出了本研究方法的消融实验结果，用于分析各模块对奶山羊爬跨行为时序动作定位性能及模型效率的影响。表中FPS为在相同硬件环境和推理配置下统计得到的单卡推理速度。以基准AdaTAD为起点，其平均mAP为76.72%，推理速度为66.19帧/s，其中可训练参数量为27.703 M，总参数量为49.583 M。

在AdaTAD基础上提出MSMA后，平均mAP提升至81.00%（+4.28个百分点），tIoU=0.7提升至69.47%，表明MSMA能增强对局部运动模式与多尺度时序信息的建模能力。同时，模型总参数量由49.583 M略增至49.815 M、可训练参数量由27.703 M增至27.936 M，推理速度下降至58.51 帧/s，说明该模块带来可接受的计算开销。

仅引入VPT机制时，模型平均mAP达到79.64%，tIoU=0.3达到91.56%，且推理速度略有提升（66.80帧/s）；对应可训练参数量仅增加0.005 M，表明VPT在较小参数增量下即可有效引导模型关注与爬跨行为相关的关键时序区域。

当同时增加MSMA与VPT后，模型取得最佳综合性能，平均mAP达到81.72%，相较基准AdaTAD提升5.00个百分点。在效率方面，该配置的推理速度为65.78 帧/s，与基准（66.19 帧/s）基本一致；可训练参数规模为27.941 M，相较基准模型仅增加0.238 M，整体开销保持在相对可控的范围内。可以看出本研究提出的方法在推理阶段模型仍保持稳定的运行速度，在精度提升的同时未引入显著的推理效率下降。

此外，MSMA与VPT分别侧重于局部运动建模与全局语义引导，两者关注的时序信息层次不同，因此其性能增益并非简单叠加，而是在协同作用下形成互补增强。

2.3 Prompt Tokens插入层数的消融实验

为了更清晰地分析Prompt Tokens插入层数对模型整体性能的影响，并验证其在任务建模中的有效性，本研究设计了针对Prompt Tokens插入层数的消融实验。实验分别在主干网络中插入2、4、6、8、10和12层Prompt Tokens，结果如表4所示。

表4 Prompt Tokens插入层数的消融实验结果

Table 4 Ablation results on the number of prompt-token insertion layers

单位：%

层数	avg mAP	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7
6	81.72	90.46	86.70	83.47	79.11	68.85
4	77.22	88.75	85.39	77.32	72.03	62.60
2	78.00	90.70	85.41	79.13	72.74	62.00
10	79.10	92.10	87.06	79.15	72.40	64.79
8	79.12	90.12	85.12	81.45	74.20	64.74
12	80.69	91.57	85.40	82.19	77.15	67.15

当Prompt Tokens插入层数较少（如2层或4层）时，模型引导作用不足，平均mAP分别为78.00%和77.22%。随着层数增加至6层，模型取得最佳性能，平均mAP达到81.72%；当层数进一步

增加至8、10或12层时，整体平均mAP反而下降，表明过深插入可能引入冗余引导信号，削弱任务聚焦效果。

综上，实验结果表明在主干前6层引入Prompt Tokens是较合理选择。该设置既能有效增强模型对奶山羊爬跨关键动作帧的关注，又能避免过深堆叠带来的性能退化问题。

2.4 Prompt Tokens个数的消融实验

为了更清晰地分析Prompt Tokens个数对模型整体性能的影响，并验证该超参数设置在爬跨行为时序动作定位任务中的合理性，本研究在固定Prompt Tokens插入层数设置以及其余训练与测试策略保持一致的条件下，仅改变每层插入的Prompt Tokens数量，分别设置为1、2、3、4，实验结果如表5所示。

表5 Prompt Tokens个数的消融实验结果

Table 5 Ablation results on the number of Prompt Tokens

单位：%

Prompt Tokens个数	avg mAP	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7
1	80.15	89.49	87.44	80.54	75.92	67.34
2	81.72	90.46	86.70	83.47	79.11	68.85
3	80.43	91.30	87.72	80.16	74.79	68.19
4	80.35	89.73	85.46	81.84	78.31	66.42

当Prompt Tokens个数较少（如1个）时，模型的提示信息容量受限，平均mAP为80.15%。当每层设置2个Prompt Tokens时，模型取得最佳性能，平均mAP提升至81.72%，且在tIoU=0.5/0.6/0.7时均有提升，说明适量增加Prompt Tokens能够增强对关键动作片段及边界位置的判别能力。

当Prompt Tokens个数进一步增加至3个或4个时，整体平均mAP未继续提升，反而出现回落。这说明在当前数据规模与训练策略下，过多tokens可能引入冗余引导信息，从而影响模型在高重叠阈值下的精确定位与泛化表现。

综上，每层设置2个Prompt Tokens是较为合理的选择，因此本研究后续实验均采用该配置。

2.5 MSMA消融分析实验

为验证MSMA在多尺度运动建模中的有效性，本研究在其余训练与推理配置保持一致的条件下，仅改变MSMA中并行DWConv1D的分支尺度设置，比较不同分支配置对定位性能的影响。改进前Adapter仅包含单一k=3的中尺度分支；在此基础上，本研究分别引入短尺度分支k=1、长尺度分支k=5，以及短/中/长3尺度分支联合，并进一步考察加入更大尺度分支（k=7）的效果。实验结果如表6所示。

表6 MSMA不同分支尺度设置的消融实验结果

Table 6 Ablation results of different MSMA branch-scale settings

单位：%

MSMA分支尺度设置	avg mAP	tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7
k={1，3，5}(our)	81.72	90.46	86.70	83.47	79.11	68.85
k={3}(baseline)	79.64	91.56	85.21	79.73	75.18	66.53
k={1，3}	79.33	89.15	85.77	80.54	75.97	65.20
k={3，5}	78.02	90.29	84.86	79.87	73.29	61.79
k={1，3，5，7}	80.80	91.40	86.54	81.66	76.21	68.17

由表6可知，引入3尺度分支（k=1，3，5）能够显著提升整体定位精度：相较于仅含中尺度分支的基线（k=3），平均mAP提升2.08个百分点（79.64%→81.72%）。更重要的是，该提升主要体现在更严格的阈值上：tIoU=0.5/0.6/0.7分别提升3.74/3.93/2.32个百分点，说明多尺度分支对边界回归精度与高重叠条件下的定位鲁棒性具有更直接的增益。

值得注意的是，仅增加单一尺度分支并不能稳定带来收益。当仅引入短尺度分支（k=1，3）时，平均mAP为79.33%，与基线相比并未提升；当仅引入长尺度分支（k=3，5）时，平均mAP降至78.02%，且在tIoU=0.7上下降至61.79%。这表明，MSMA的性能提升并非来源于分支数量增加，而主要来源于不同时间尺度特征之间的互补建模。其中，短尺度分支有助于捕捉突发变化与局部细节，长尺度分支有助于提供更稳定的上下文约束，中尺度分支则保持局部结构建模的稳健性；三者联合后能够在中高阈值下获得更一致的性能提升。

此外，继续加入更大尺度分支（k=1，3，5，7）并未进一步提升平均mAP（80.80% < 81.72%），表明在当前数据规模与任务时序特性下，过大的时序感受野带来的额外上下文信息并不能转化为稳定收益，反而可能引入冗余聚合与优化负担。因此本研究采用k=1，3，5作为MSMA的默认分支尺度设置。

2.6 训练集内部5折稳定性评估实验

为进一步评估模型在既定数据划分方案下的性

能稳定性，并减弱单次随机划分可能带来的偶然性影响，本研究在训练集内部开展5折划分稳定性评估，并在固定外部测试集上统计模型性能的波动范围，实验结果如表7所示。

表7　固定外部测试集上的5折稳定性评估结果

Table 7　Five-fold stability evaluation results on a fixed external test set

单位：%

Fold/统计量	Avg-mAP	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7
Fold 1	77.43	88.44	84.78	79.60	73.11	61.21
Fold 2	79.00	88.26	85.07	79.81	75.76	66.10
Fold 3	80.81	89.52	86.04	80.87	77.91	69.72
Fold 4	82.19	94.40	87.96	84.77	77.58	66.23
Fold 5	80.74	90.10	85.08	81.44	76.56	70.54
Mean	80.03	90.14	85.79	81.30	76.18	66.76
Std.	1.84	2.50	1.31	2.08	1.92	3.69

具体而言，该实验仅在65段训练视频上进行：首先按视频ID对训练视频进行随机打乱，并划分为5个互斥子集（每折约13段）。在第k折实验中，选取其中4个子集（约52段视频）作为训练集，剩余1个子集（约13段视频）作为验证集，用于训练过程中的模型选择；训练完成后，将该折模型在固定的77段测试视频上进行1次评估。为保证实验可复现性，5折划分在实验开始前一次性生成并保存为文本文件，后续所有折均严格复用该划分；外部测试集在整个过程中不参与任何形式的参数选择与模型调优，仅用于最终性能评估与波动统计。

本研究方法在固定测试集上的Avg-mAP均值为80.03%，标准差为1.84，表明在不同训练/验证划分下模型性能波动较小。各tIoU阈值下的mAP随阈值增大而下降，其中高阈值处标准差相对更大，说明在严格边界条件下模型评估对边界偏移更为敏感。

此外，表7的5折统计均值较表2的单次二分划分结果略低，这主要与每折训练视频数量进一步减少有关。总体来看，5折稳定性评估通过报告“均值±标准差”对性能波动进行了定量刻画，增强了实验结论的稳健性。

2.7　全数据集5次5折交叉验证实验

尽管第2.6节已对固定外部测试集上的性能波动进行了统计分析，但该实验仍基于单一训练/测试划分方案。为进一步从统计角度评估模型在整个数据集上的稳定性与泛化表现，在全部142段未修剪视频上开展5次5折交叉验证实验，并统计不同重复实验下模型性能的均值与标准差，结果如表8所示。

表8　全数据集5次5折交叉验证实验结果

Table 8　Results of 5×5-fold cross-validation on the entire dataset

单位：%

统计量	Avg-mAP	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7
Repeat1	80.83±2.63	90.79±1.80	86.56±1.85	82.23±2.32	76.85±3.34	67.70±3.91
Repeat2	80.92±2.08	90.74±1.47	86.56±1.44	82.25±1.71	76.93±2.73	68.13±3.17
Repeat3	80.91±2.35	90.73±1.64	86.53±1.73	82.27±2.04	76.97±3.00	68.07±3.40
Repeat4	80.76±2.76	90.62±1.96	86.39±1.99	82.17±2.35	76.81±3.37	67.79±4.19
Repeat5	80.92±2.28	90.71±1.53	86.63±1.71	82.23±1.95	76.89±2.89	68.13±3.42
Overall	80.87±2.22	90.72±1.54	86.53±1.60	82.23±1.91	76.89±2.81	67.97±3.33

在每一次重复实验中，首先将全部142段视频随机划分为5个互斥子集，每次取其中4个子集作为训练数据、1个子集作为测试数据进行评估；完成5折训练与测试后，统计该次重复实验的性能均值与标准差。为降低随机初始化与数据划分带来的偶然性影响，上述过程在不同随机种子下重复5次，并最终对25次实验结果进行汇总统计。

表8表明，5次重复实验的Avg-mAP均值稳

定在80.76%~80.92%，整体平均性能为80.87%±2.22%，说明模型在不同数据划分与随机初始化条件下仍能够保持较为稳定的检测性能。各tIoU阈值下的结果同样呈现出随着阈值提高而逐渐下降的趋势。

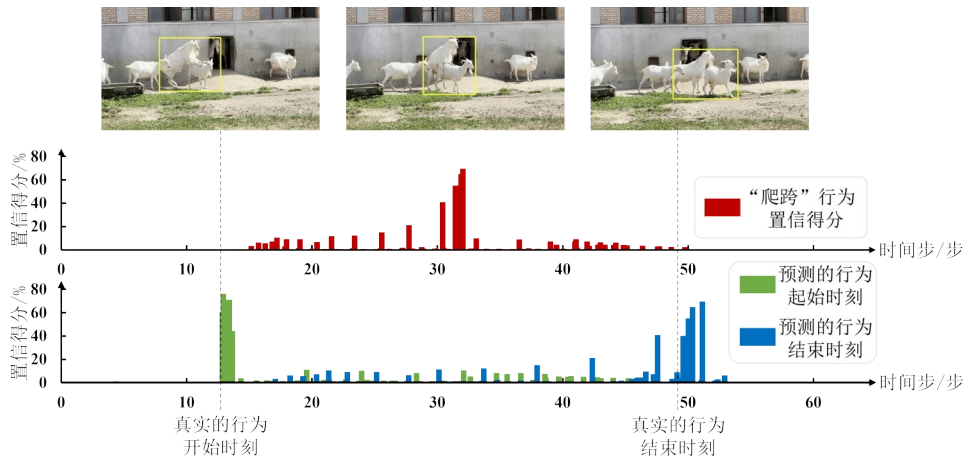
从标准差角度看，低阈值条件下模型结果较为稳定，而在更严格的边界约束条件下波动有所增加。这一现象与表7中的规律一致，说明当评估对时间边界精度要求更高时，模型性能对训练样本划分与行为边界差异更加敏感。

进一步比较主实验（表2）与交叉验证结果可以发现，5×5交叉验证的整体平均性能（80.87%）与主实验的单次二分划分结果（81.72%）基本一致，仅存在约0.85个百分点的差异。这种差异主要来源于交叉验证过程中不同训练/测试划分带来的统计波动，而非模型性能的不稳定。通过对25次实验结果进行汇总统计，5×5交叉验证能够在更大程度上降低单次数据划分带来的偶然性影响，从而从统计意义上验证了本研究方法在奶山羊爬跨行为时序动作定位任务中的稳定性与可靠性。

综合2.6节与本节实验结果可以看出，无论是在固定外部测试集条件下的训练集5折稳定性评估，还是在全数据集范围内的多次交叉验证实验中，本研究方法均表现出较为稳定的性能水平，说明所提出的VPT与MSMA结构在不同数据划分条件下均能够稳定提升模型对奶山羊爬跨行为的时序定位能力。

2.8 可视化结果分析

为进一步直观验证所提出模型在奶山羊爬跨行为时序动作定位任务中的有效性，本研究选取典型室外与室内养殖环境下的视频片段进行可视化分析，结果分别如图5和图6所示。所选示例涵盖无遮挡与遮挡、不同拍摄视角及环境条件，具有代表性，并非仅选取预测效果最优的视频片段。



注：上方为关键帧示例；中间红色柱为分类分支对“爬跨”类别的时序置信度（Sigmoid后）；下方绿色/蓝色柱为回归分支在各时间位置预测的起始/结束偏移量解码为边界位置后，在时间轴上统计/加权汇总得到的起始/结束边界响应强度（用于可视化展示，并非额外输出的概率分布）。虚线为人工标注的起止时间。

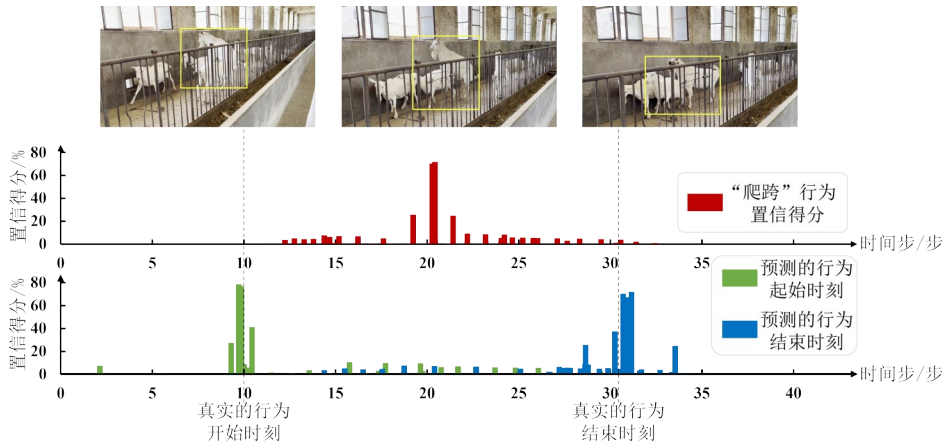
图5 模型在奶山羊室外爬跨行为视频上的预测结果示例

Fig. 5 Examples of prediction results of the model on videos of outdoor mounting behavior in dairy goats

从图5可以看出，在室外自然光照条件下，尽管背景环境较为复杂、个体间存在一定程度的交互干扰，模型仍能够在行为发生前后准确捕捉到爬跨行为的时序变化趋势。红色置信度在目标行为区间内形成清晰主峰，绿色与蓝色峰值分别分布于主峰区间的前后边界位置，且与人工标注的起止时间基本一致，表明模型在开放场景下具备较强的时序建模能力和边界判别能力。

图6展示了模型在室内圈舍环境中的定位结果。相比室外场景，室内背景结构相对固定，但存在光照不均、个体遮挡频繁等问题。从可视化结果可以观察到，模型同样能够在爬跨行为发生区域内产生稳定且集中的高置信度响应，起始与结束边界的响应峰值整体分布于真实边界附近，仅在个别时间点出现轻微偏移，说明所提出方法在受限空间和高密度个体场景下仍具有较好的鲁棒性。

需要指出的是，由于奶山羊爬跨行为持续时间较短、动作幅度相对细微，模型在部分非目标背景帧上可能产生低幅度响应。然而，这类响应明显低于目标行为区域内的主峰值，且未形成连续高响应



注：上方为关键帧示例；中间红色柱为分类分支对“爬跨”类别的时序置信度（Sigmoid 后）；下方绿色/蓝色柱为回归分支在各时间位置预测的起始/结束偏移量解码为边界位置后，在时间轴上统计/加权汇总得到的起始/结束边界响应强度（用于可视化展示，并非额外输出的概率分布）。虚线为人工标注的起止时间。

图 6　模型在奶山羊室内爬跨行为视频上的预测结果示例

Fig. 6　Examples of prediction results of the model on videos of indoor mounting behavior in dairy goats

区间，因此不会对最终的时序定位结果产生实质性干扰，体现了模型对关键动作帧与背景帧的有效区分能力。

3　结　论

1）针对奶山羊爬跨行为持续时间短、动作幅度小且易受遮挡等问题，本研究在 AdaTAD 端到端时序动作定位框架的基础上，提出融合 VPT 机制与 MSMA 的改进模型。VPT 机制通过在主干网络中引入少量可学习 Prompt Tokens，实现对关键帧的任务引导与特征聚焦；MSMA 模块采用并行多尺度卷积结构自适应捕捉不同时间尺度下的运动模式，提升短时微动作与长时宏观动作的协同建模能力，有效增强了模型的时空特征表达能力。

2）在所构建的奶山羊爬跨行为数据集上的实验结果表明，本方法在 tIoU 阈值［0.3∶0.1∶0.7］下的平均 mAP 达到 81.72%，在高阈值 tIoU=0.7 时的 mAP 为 68.85%，分别较 TadTR、VSGN、AFSD、ActionFormer、TriDet、DyFADet、AdaTAD 和 Re2TAL 等主流模型提升 1.20～38.82 个百分点。提出 MSMA 模块后平均 mAP 提升 4.28 个百分点，进一步结合 VPT 机制后整体性能最优，验证了两者在特征聚焦与多尺度建模方面的协同增益。

3）通过可视化结果分析可以看出，所提出模型在大多数复杂养殖场景下能较为准确定位奶山羊爬跨行为的发生区间及其起止时间，但在严重遮挡及动作姿态高度相似等极端条件下，模型仍可能出现边界定位偏差或漏检现象。这些分析结果从侧面揭示了当前模型在真实养殖环境中的适用边界，也为后续改进提供了明确方向。

综上，本研究验证了端到端时序动作定位方法在特定养殖场景下奶山羊爬跨行为监测任务中的应用可行性，为智能化繁殖行为感知提供了基础技术支撑。从实际生产流程角度看，所提出的方法可作为繁殖管理中行为感知层的关键组成模块，未来可进一步与羊只个体身份识别相结合，形成面向个体的爬跨行为时序记录，为精准配种决策、漏配补救提供数据基础。未来工作将进一步引入多养殖场、多环境条件下的数据，对模型的跨场景泛化能力与长期稳定性进行系统评估，并结合复杂遮挡感知建模、短时行为精细表征及多模态信息融合等方向，持续提升模型在真实养殖场连续监测场景中的适用性与鲁棒性。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] 王引, 张秋桐, 樊平. 养羊业发展趋势及针对性防疫措施[J]. 中国动物保健, 2024, 26(4): 87-88.

[2] 万玉辉, 陈新文, 左晓佳, 等. 基于改进 YOLOv11 的羊只检测方法研究[J]. 草食家畜, 2026(2): 64-70.
WAN Y H, CHEN X W, ZUO X J. Research on a sheep detection method based on improved YOLOv11 module[J]. Grass-Feeding Livestock, 2026(2): 64-70.

[3] ENDO N, RAHAYU L P, ARAKAWA T, et al. Video tracking analysis of behavioral patterns during estrus in goats[J]. The Journal of Reproduction and Development, 2016, 62(1): 115-119.

[4] KOLAREVIC J, AAS-HANSEN Ø, ESPMARK Å, et al. The use of acoustic acceleration transmitter tags for monitoring of Atlantic salmon swimming activity in recirculating aquaculture sys-

tems (RAS)[J]. Aquacultural Engineering, 2016, 72: 30-39.
[5] NASIRAHMADI A, EDWARDS S A, STURM B. Implementation of machine vision for detecting behaviour of cattle and pigs [J]. Livestock Science, 2017, 202: 25-38.
[6] WANG J, SI Y F, WANG J P, et al. Discrimination strategy using machine learning technique for oestrus detection in dairy cows by a dual-channel-based acoustic tag[C]// Computers and Electronics in Agriculture. New York, USA: ACM, 2023.
[7] WANG Z, HUA Z X, WEN Y C, et al. E-YOLO: Recognition of estrus cow based on improved YOLOv8n model[J]. Expert Systems with Applications, 2024, 238: 122212.
[8] WANG R, GAO R H, LI Q F, et al. A lightweight cow mounting behavior recognition system based on improved YOLOv5s[J]. Scientific Reports, 2023, 13: 17418.
[9] SHI J R, CHEN X W, ZHANG Y L, et al. Detection of estrous ewes' tail-wagging behavior in group-housed environments using Temporal-Boost 3D convolution[J]. Computers and Electronics in Agriculture, 2025, 234: 110283.
[10] DUAN Y P, YANG Y Z, CAO Y, et al. A multimodal deep learning network for precise detection of estrus and pseudo-estrus in sows[J]. Smart Agricultural Technology, 2025, 12: 101279.
[11] CAO Y, YIN Z, DUAN Y P, et al. Research on improved sound recognition model for oestrus detection in sows[J]. Computers and Electronics in Agriculture, 2025, 234: 109975.
[12] WANG Z, DENG H X, ZHANG S J, et al. Detection and tracking of oestrus dairy cows based on improved YOLOv8n and TransT models[J]. Biosystems Engineering, 2025, 252: 61-76.
[13] 张志勇, 曹姗姗, 孔繁涛, 等. 母牛发情精准感知与智能鉴定技术研究进展、问题与挑战[J]. 智慧农业(中英文), 2025, 7(3): 48-68.
ZHANG Z Y, CAO S S, KONG F T, et al. Advances, problems and challenges of precise estrus perception and intelligent identification technology for cows[J]. Smart Agriculture, 2025, 7(3): 48-68.
[14] LIU X L, BAI S, BAI X. An empirical study of end-to-end temporal action detection[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2022: 19978-19987.
[15] CHEN G, ZHENG Y D, WANG L M, et al. DCAN: Improving temporal action detection *via* dual context aggregation [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California: AAAI Press, 2022: 248-257.
[16] ZHU Z X, TANG W, WANG L, et al. Enriching local and global contexts for temporal action localization[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, New Jersey, USA: IEEE, 2021: 13496-13505.
[17] XU M M, ZHAO C, ROJAS D S, et al. G-TAD: Sub-graph localization for temporal action detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2020: 10153-10162.
[18] ZENG Y S, ZHONG Y J, FENG C J, et al. UniMD: Towards unifying moment retrieval and Temporal action detection[C]// Computer Vision-ECCV 2024. Cham, Germany: Springer, 2025: 286-304.
[19] LIU S M, ZHANG C L, ZHAO C, et al. End-to-end temporal action detection with 1B parameters across 1000 frames[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2024: 18591-18601.
[20] JIA M L, TANG L M, CHEN B C, et al. Visual prompt tuning [C]// Computer Vision-ECCV 2022. Cham, Germany: Springer, 2022: 709-727.
[21] LIU X L, WANG Q M, HU Y, et al. End-to-end temporal action detection with transformer[J]. IEEE Transactions on Image Processing, 2022, 31: 5427-5441.
[22] ZHAO C, THABET A, GHANEM B. Video self-stitching graph network for temporal action localization[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, New Jersey, USA: IEEE, 2021: 13638-13647.
[23] LIN C M, XU C M, LUO D H, et al. Learning salient boundary feature for anchor-free temporal action localization[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2021: 3319-3328.
[24] ZHANG C L, WU J X, LI Y. ActionFormer: Localizing moments of Actions with Transformers[C]// Computer Vision – ECCV 2022. Cham, Germany: Springer, 2022: 492-510.
[25] SHI D F, ZHONG Y J, CAO Q, et al. TriDet: Temporal action detection with relative boundary modeling[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2023: 18857-18866.
[26] YANG L, ZHENG Z W, HAN Y Z, et al. DyFADet: Dynamic feature aggregation for Temporal action detection[C]// Computer Vision – ECCV 2024. Cham, Germany: Springer, 2025: 305-322.
[27] ZHAO C, LIU S M, MANGALAM K, et al. Re2TAL: Rewiring pretrained video backbones for reversible temporal action localization[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2023: 10637-10647.

Temporal Action Localization of Mounting Behavior in Dairy Goats Based on Improved AdaTAD

WANG Jiayuan[1], LI Qitong[1], LUO Yuantao[1], YANG Shuqin[2], WANG Zhenhua[1], NING Jifeng[1*], WANG Meili[1]

(1. *College of Information Engineering*, *Northwest A & F University*, *Yangling* 712100, *China*;
2. *College of Mechanical and Electronic Engineering*, *Northwest A & F University*, *Yangling* 712100, *China*)

Abstract: [Objective] Accurate temporal localization of mounting behavior in dairy goats is important for intelligent reproductive management, as event frequency, onset time, and duration provide useful evidence for heat monitoring and mating decisions. Unlike simple behavior recognition, temporal localization in untrimmed videos enables fine-grained, time-resolved records for practical farm use. However, real-world mounting behavior is usually brief and sporadic, with few informative frames in long video streams. Moreover, weak discrimination from similar non-target interactions, together with occlusion, viewpoint variation, and background motion, often

degrades boundary-aware representation learning and leads to unstable start-end localization. To enhance localization accuracy and stability while maintaining practical efficiency for deployment, an improved AdaTAD-based end-to-end temporal action localization approach was proposed for mounting behavior in dairy goats. **[Methods]** The proposed approach adopted AdaTAD as the baseline end-to-end temporal action localization framework and introduced two complementary improvements, explicit key-frame guidance and multi-scale motion modelling, while retaining the original detection head and post-processing pipeline for generating temporal action instances. First, visual prompt tuning (VPT) was incorporated to provide task-conditioned guidance to backbone feature extraction in a parameter-efficient manner. Specifically, a small number of learnable prompt tokens were inserted into the Transformer backbone with backbone parameters frozen. Through multi-head attention interactions between prompt tokens and patch tokens, the prompts steered attention towards mounting-relevant temporal regions, strengthened feature responses at critical frames and in boundary neighbourhoods, and improved the separability between brief target segments and abundant background frames. Second, a multi-scale motion adapter (MSMA) was introduced to model motion patterns at different temporal scales and improve robustness to diverse scene dynamics. MSMA employed parallel multi-scale temporal depthwise separable convolution branches to capture short-, mid-, and longer-range temporal variations, enhancing representations of subtle short-duration micro-actions as well as relatively complete action processes. Residual connections and nonlinear mappings further stabilised feature injection and gradient propagation, enabling multi-scale dynamics to be integrated into backbone features with limited additional optimisation burden. Overall, VPT focused on boundary-relevant attention guidance, whereas MSMA emphasised multi-scale temporal dynamics modelling; Together, they formed a complementary design within the end-to-end localization pipeline. **[Results and Discussions]** Comparative experiments showed that the proposed method achieves an average mAP (mean Average Precision@[0.3:0.1:0.7]) of 81.72%, improving upon the baseline AdaTAD by 5.00 percentage points, indicating that incorporating VPT and MSMA enhanced overall localization performance. At a temporal Intersection over Union (tIoU) threshold of 0.7, the proposed method attained 68.85%, exceeding AdaTAD by 4.06 percentage points, demonstrating that the performance gain was preserved under stricter temporal boundary-consistency criteria. Further comparisons with representative approaches, including TadTR, VSGN, AFSD, ActionFormer, TriDet, DyFADet, and Re2TAL, showed average mAP improvements of 38.82, 33.83, 25.29, 4.09, 2.83, 1.20, and 6.06 percentage points, respectively, demonstrating stronger overall competitiveness. In terms of efficiency, the model ran at 65.78 f/s with 27.941 million trainable parameters, indicating that the accuracy gains were achieved while maintaining a relatively low parameter overhead and practical runtime efficiency. Overall, task-guided prompting and multi-scale temporal modelling improved key temporal feature representations with limited parameter increments, thereby benefiting localization of short, sporadic behaviors. **[Conclusions]** By combining VPT for boundary-relevant attention guidance with an MSMA for multi-scale temporal dynamics modelling, the proposed AdaTAD-based end-to-end temporal action localization method improves localization accuracy and maintains stable advantages under stricter boundary-consistency requirements, while preserving practical inference efficiency. The method could provide temporal information for reproductive behavior monitoring and decision support, and offer a feasible basis for building individual-level, time-resolved management systems in real farming environments.

Key words: mounting behavior; temporal action localization; AdaTAD; precision livestock farming; multi-scale motion modeling; end-to-end model; smart breeding

Foundation items: National Key Research and Development Program of China (2022YFD1300200); Shaanxi Qinchuangyuan High-level Innovation and Entrepreneurship Talent Program (QCYRCXM-2022-359)

First author: WANG Jiayuan, E-mail: yolo_wjy@nwafu.edu.cn
***Corresponding author:** NING Jifeng, E-mail: njf@nwsuaf.edu.cn