arXiv:2301.00482v2 [cs.HC] 7 Jan 2023

FEVA：快速事件视频标注工具

Snehesh Shrestha, William Sentosatio, Huiashu Peng, Cornelia Fermüller, Yiannis Aloimonos马里兰大学帕克分校，马里兰州帕克市 20740snehesh@umd.edu

摘要

视频标注是计算机科学和社会科学领域都至关重要的过程。许多视频标注工具（VATs）提供了丰富的功能，使视频标注成为可能。我们对 59 种以上的视频标注工具进行了广泛调查，并访谈了跨学科研究人员，以评估视频标注工具的易用性。研究结果表明，目前大多数视频标注工具的用户界面令人眼花缭乱，交互技术不佳，功能也难以理解。这些问题往往会导致标注时间更长、标签不一致以及用户疲劳。我们推出了 FEVA，这是一款视频标注工具，采用精简的交互技术和动态界面，让标注任务变得轻松快捷。FEVA 注重速度、准确性和简洁性，使标注快速、一致且直观。例如，标注人员只需按下单个按键，即可控制视频的播放速度和方向，并实时标记标签的起始时间和结束时间。在我们的用户研究中，FEVA 用户平均所需的交互操作比最受欢迎的标注工具（Advene、ANVIL、ELAN、VIA 和 VIAN）少 36%。参与者（N=32）认为 FEVA 更直观，且所需的心理需求更低。代码和演示可在 http://www.snehesh.com/feva 获取。

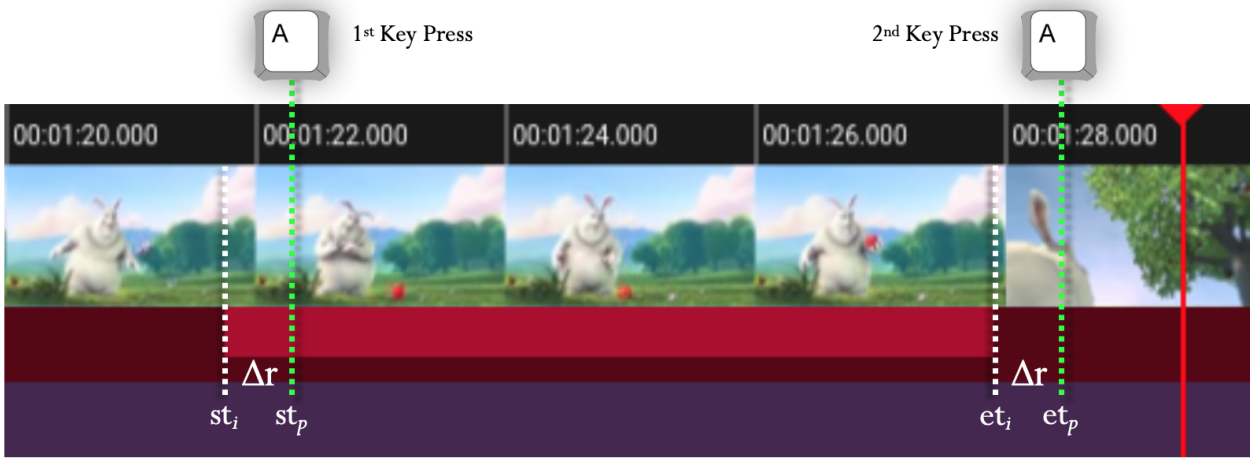


图1：“速度标注”允许您通过分别使用一次按键标记开始时间（st_p）和结束时间（et_p），实时创建标注（红色矩形）。FEVA 会根据您的反应时间（Δr）自动调整该事件标注的时间，从而为您提供最精确的预期时间（st_i 和 et_i）。

关键词 视频标注工具 · 用户界面设计 · 用户交互

1 引言

社会科学家需要对录像访谈和实验中的对话与行为进行编码，而这些内容很快就会累积成数小时的录像 [1, 2, 3]。计算机科学家需要包含适当标注的真实标注、用于机器学习的数据集，其中的视频片段总时长可达数百小时 [4, 5, 6], , 甚至数千小时 [7]。因此，视频标注既耗时又繁琐。

现有的视频标注工具 [8, 9, 10, 11, 12, 13] 提供了多种适用于视频标注活动的功能。然而，由于学习曲线陡峭、功能复杂、界面令人应接不暇，以及交互技术欠佳等问题，它们往往无法满足研究人员的需求，进而导致标注时间延长、结果不一致和用户疲劳。

例如，为了分析足球比赛，研究人员会标注球员控球、踢球和助攻。使用每次都需要暂停并让用户为标注命名的工具、必须通过鼠标上下文菜单选项进行标注的工具，以及不允许标注重叠的工具，会使这一任务变得极其繁琐且耗时。相比之下，手动编码更加直接。标注人员可以降低视频播放速度，使用点击器计数或秒表 [14] 实时记录时间戳，然后在完成后将其输入电子表格。因此，许多标注人员仍然采用手动编码，并依赖电子表格 [15]。当需要处理数小时此类视频素材时，计算机科学家通常会将这些标注任务外包或众包 [16, 17, 18, 19]。这可能引发对隐私和可靠性的担忧 [20]。为了解决这一问题，研究人员会对人脸进行模糊处理，以隐藏参与者身份。然而，这可能会损害情绪判断的质量，导致结果不一致。

本文旨在设计一款视频标注工具，使标注任务变得轻松快捷。我们访谈了来自不同学科的研究人员，以了解视频标注所采用的标准做法、工作流程和工具。具体而言，我们访谈了来自5个领域的13名研究人员（神经科学、行为心理学、电影研究和计算机科学）。研究人员对现有视频标注工具表示保留，因此会避开这些工具。我们还调查了59款视频标注工具（列表见附录）。我们根据亲身体验对其主要功能和界面设计选择进行了分类，并分析了无法访问的工具的视频教程。根据访谈和调查结果，我们提出了五项有助于视频标注活动的设计标准，详见第4节：

- D1. 根据操作工作流程组织空间
- D2. 简化高频操作
- D3. 尽可能采用算法支持
- D4. 沿用有效做法，并重新设计无效之处
- D5. 保持灵活性

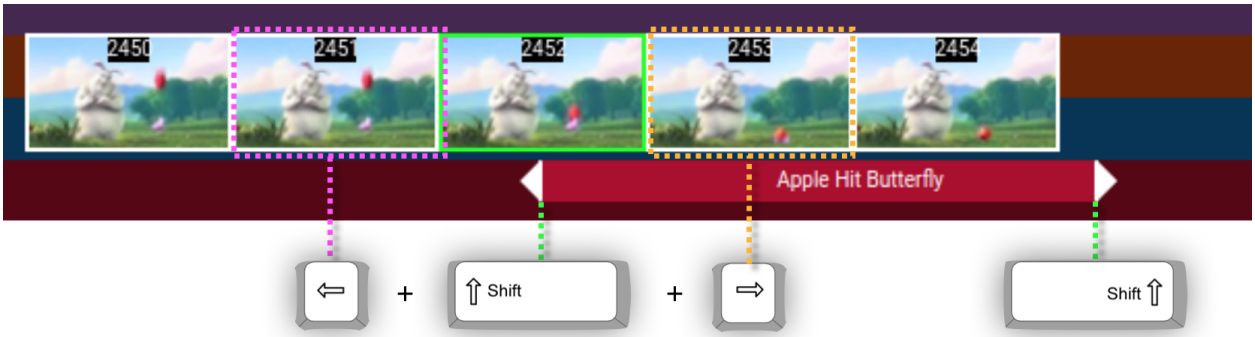


图2：按下 Shift键可启用微调功能。左侧和右侧 Shift键分别对应标签的开始时间和结束时间。按住 Shift 键的同时按下箭头键，可将标签的开始或结束时间调整一帧。

这些标准为快速事件视频标注（FEVA）的设计提供了依据。FEVA 是一款视频标注工具，采用简化的交互技术和动态界面，使标注任务变得轻松快捷。简化用户界面以及实时标注与反应时间调整等功能（图1），再加上精确的微调机制（图2），能够帮助标注人员比任何视频标注工具更快地创建大量准确的标签。

为评估 FEVA，我们开展了两项比较研究。在第一项研究中，我们比较了用户使用 FEVA 与另外五种基于事件的视频标注工具完成任务所需的输入次数 [8, 9, 10, 12, 13]。如表3所示，平均而言，与竞争性视频标注工具相比，FEVA 完成相同任务所需的输入次数减少了 36.0%。在第二项研究中，我们要求 34 名参与者以随机顺序使用其中一种视频标注工具进行标注。我们发现，无论参与者的背景如何，他们认为 FEVA 更直观的比例为 88%，认为其更易于使用的比例为 91%。用户还评价 FEVA 的心理需求低 46%（$p < 0.00003$），挫折感少 62%（$p < 0.00147$），体力需求低 41%（$p < 0.00187$），所需努力程度低 34%（$p < 0.00324$）。

总之，本文的贡献如下：

- 通过访谈、工具调查和试点测试，全面了解**现有的视频标注工具**。
- 一份针对视频标注工具**设计**的**标准**清单
- **FEVA**，一款基于事件的视频标注工具，可帮助您更快速、更准确地进行标注。
- 一项比较研究和一项**用户研究** 表明，FEVA 更直观、所需努力程度更低，同时还能生成一致性更高、准确度更高的标注。

2 相关工作

本节将介绍基于事件的标注、我们所基于的工具、标注过程中采用的工作流程，以及用户界面和相关的交互技术。为了了解标注工作流程和目标，我们采访了研究人员，并调查了有关完成所需标注步骤的文献。我们发现，标注人员主要分为两类：一类借助视频标注工具进行标注，另一类则依赖启发式方法。

2.1 基于事件的标注

视频标注是标记感兴趣区域（ROI）的过程，分为 a）空间标注（对象标注，OA）或 b）时间标注（事件标注，EA）。OA 标记的是单个时间点，称为基于帧的标注。EA 指事件发生的时间段。计算机视觉领域广泛采用 OA 进行目标检测和识别，并使用边界框 [21, 22, 23, 24], 点 [25, 12], 或多边形（即分割） [26, 27, 28] 或掩膜 [25, 29, 23] 来标记对象。虽然一些视频标注工具也会 [30, 31]随时间 [12, 25, 22, 24], 追踪物体，但这与 EA 不同。EA 关注场景中发生了什么、物体在做什么，或物体遭遇了什么 [32, 33]，而不是物体本身。因此需要标记事件的开始和结束时间。EA 包括行为、互动、情绪反应、言语和动作 [34, 35, 36]。其应用范围涵盖多个学科：从计算机科学中的活动识别 [37, 12, 38, 39, 40, 41], 、心理学中的行为分析 [11, 42, 35], , 到新闻学中沿时间线追踪并呈现故事 [43]。然而，事件标注可能颇具挑战，因为时间导航本身较为复杂，需要能够在连续变化的事件中准确标记所需时间，还需要利用工具的现有功能来辅助标注。因此，开发 FEVA 的重点是简化工作流程并优化事件标注步骤，使标注更加轻松、高效。

2.2 启发式工作流程

有些研究人员不愿使用专门的视频标注工具。例如，有些研究人员认为，使用点击器记录某些事件发生的时间更容易。这种方法容易出错，可靠性较低，而且不便于日后查阅记录。尽管有些视频标注工具可以通过按键操作提供类似功能 [12] ，研究人员仍对使用视频标注工具有所顾虑，担心初始设置耗时，以及新编码人员需要反复学习。在另一项研究中 [14], , 研究助理（RA）使用秒表测量“从按下确定按钮到设备发出提示音表示RR计数完成所花费的时间”。这些点击器计数或秒表计时结果要么用笔和纸手工记录，要么录入电子表格。另一种工作流程中，一位人际关系研究者介绍说，他们让研究助理观看两人互动的视频，并回答研究人员提出的问题。例如，多名研究助理关注整个互动过程或视频中的某个特定人物，并用李克特量表评估一方在另一方说话时的专注程度。这些结果要么记录在纸上，要么通过 Qualtrics 或 Google 表单在线填写。

2.3 视频标注工具工作流程

使用视频注释工具的主要工作流程与简单编码所采用的启发式方法并无不同。主要区别在于学习曲线更陡，设置过程也更长。然而，随着编码变得更加复杂，启发式方法

随着标注人员增多，启发式方法在速度方面的收益会逐渐递减。标注过程会慢得多，通常需要人工完成许多工具本可以简化的步骤。不过，并非所有视频标注工具都一样。不同视频标注工具的工作流设计、功能的屏幕布局、交互步骤和技术都存在显著差异。

计算机科学家和工程师为计算机视觉、机器学习和机器人学研究及应用设计了大多数视频标注工具 [12, 35, 4, 16, 7]。不过，也有少数工具专为其他领域设计并应用于这些领域，例如电影研究 [13, 44] ，视频标注工具可用于角色分析和场景分析。新闻记者使用视频标注工具标注并同步来自多个来源的事件，以呈现连贯的故事 [43]。在体育领域，团队会标注比赛和重要事件，并通过回看进行研究，服务于各自的队伍 [45, 46]。尽管大多数工具各有侧重，标注的核心是标记事件是否存在，并确定感兴趣事件的时间范围。为此，事件标注是后续工作得以更轻松开展、甚至在某些情况下成为可能的基础。因此，准确标注至关重要。支持同步多摄像机视图、在时间轴上对事件进行微观和宏观可视化，以及便捷地搜索和回看视频等功能，使 FEVA 成为这些领域的理想选择。

2.3.1 布局设计：功能与屏幕空间使用的平衡

有些工具具备大量功能，工作流程复杂，可自定义和使用的功能组合也很多，因此用户界面中充斥着小窗口、按钮以及层层菜单项，用户需要通过这些元素来访问相应功能。这些工具需要花费更长时间才能学会并适应；然而，一旦掌握，它们可能非常强大。由于功能众多，例如媒体播放器控件、标注控件和可视化，有限的屏幕空间可能会带来很大挑战。因此，有些工具会将这些功能归入较小的窗口 [13, 9, 8], ，通过多级菜单 [9, 10], 或不同的操作模式 [12, 10, 13]对其进行组织。虽然这有助于整理功能，但访问起来并不容易，用户需要大量练习才能记住使用软件所需的各种步骤。与此同时，也有一些简单得多的工具，只服务于非常特定的用途 [46, 47, 44, 48, 49], ，其布局简洁，并且能够充分利用屏幕空间来实现所提供的功能。这些工具易于学习和开始使用。然而，它们的功能也受到限制。我们的访谈发现，大多数研究人员和标注人员的需求处于两者之间。复杂的 VAT 中，大多数功能从未被使用；而简单的 VAT 又限制了用户完成其特定用途之外的任务，因此他们需要将其与其他工具配合使用，以弥补差距并满足完整需求。我们设计 FEVA 的动机，是创建一种易于导航和使用的工具，同时具备更复杂工具所提供的功能，又不会复杂到让用户必须参加课程才能使用。

2.4 采用

VIA [12] 的键盘快捷键直观且实用，尤其是根据上下文弹出的提示；大多数工具都缺少这一功能，而 FEVA 也采用了这一设计。FEVA 使用了媒体播放器中已成为标准的常用快捷键，例如按空格键播放或暂停，按 ctrl+Z 撤销，按 ctrl+Y 重做等。VIAN [13] 是唯一一款无需先选中标签，就能用鼠标拖动标签的工具，FEVA 也采用了相同的方式。Advene [9]、ELAN [10] 和 VIA [12] 提供了其他方式来显示或执行播放选项，例如连续模式。虽然这些方式在某些情况下很有用，但大多数标注人员都使用默认方式，没有进行更改，这可能会令人困惑。有些工具 [12] 允许用户在影片播放时创建标签，但只记录开始时间，并采用固定时长。遗憾的是，大多数用户之后还需要返回标签并重新调整。FEVA 改进了这种互动方式，允许用户再次按键来标记端点。

3 了解最先进的视频注释工具及其比较

3.1 目标用户

为了了解不同学科的研究人员是否会标注数据、这涉及哪些工作，以及工作流程是什么样的，我们采访了 3 位神经科学研究人员、3 位行为心理学研究人员、2 位电影研究教师和 5 位计算机科学家。访谈采用半结构化形式，旨在回答以下 3 个问题：

- **研究问题1**：其涉及人类研究的研究性质、收集的数据类型，以及是否涉及视频 o.
- **研究问题2**：数据收集过程中的工作流程、这些数据的后处理以及代码生成。
- **研究问题3**：视频标注的结构与工作流程，以及标注之后的用途。

访谈结束后，我们对回答进行了汇总，并对技术挑战进行了编码。访谈洞见（II）如下：

- **II1**。主要有三种时间标注。1）用于标记某些事件是否出现的二元标签。例如，研究人员感兴趣的是统计“一个人触摸自己

脸部的次数，以此作为参与者有多紧张的一项指标。” 2）范围标签用于标记感兴趣事件的开始和结束。“参与者会为视频中的特定时刻添加标注。例如，一对伴侣可能会进行互动，之后研究人员让每位伴侣观看视频，并在某个特定想法或感受出现时进行标注。他们会对两位参与者都这样做，以观察想法、感受、目标等方面的趋同与不一致。” 3）某些类型的标签（例如情绪或场景标签）持续的时间比动作标签更长。“例如，我们可能想评估某人对其伴侣的体贴或关怀程度。因此，研究助理可能会观看一整段互动，专注于视频中的某一位参与者，然后回答一个关于伴侣说话时此人有多专注的问题。”大多数视频标注工具对这类标签的支持都很差，因此研究人员会采用变通办法，例如创建多个标签轨道，每个轨道专门用于记录一种特定反应。

- **II2**：数据标注所需的时间、努力程度和成本都高得惊人。因此，任何能够改进标注流程、使其更快速、更可靠的系统，都能带来巨大的益处。“我们花了很多研究助理的工时来做这些标注。如果有一种工具哪怕能节省一小时，我肯定会用它。”
- **II3**：即使有非常明确的编码规则，已标注数据的时间精度往往仍然较低，因此放宽了标注一致性规则。“许多研究助理都会查看并标注这些视频。每位研究助理对某个事件发生时间的判断都会有些许差异。”
- **II4**：在标注过程和分析阶段，协作和共享标注内容都很繁琐，需要团队之间经过多个步骤才能保持同步。“.....如果学生同时在标注同一个片段，就需要刷新页面才能看到同学写了什么。这对我来说也是个问题，因为我喜欢在他们工作时添加评论（鼓励他们详细阐述某些观点，或探索新的观点）。”
- **II5**：当前的视频注释工具界面不佳，不便于研究人员探索、分析和搜索已标注数据。
- **II6**：众包或人工智能生成的标注往往需要大量审核，因此研究人员让研究助理完成标注反而更容易、更快捷。“所以我使用了 [在线自动语音识别工具]。即使使用高级版，我还是得逐一编辑所有内容。动作标签和逐字语音标签的起止时间必须相互对应。所以对我来说，手动创建标签其实更容易。”

本文聚焦于 II2、II3 和 II5，这些内容有助于制定第4节中的设计决策 .

3.2 视频标注工具的规格比较

我们根据附录B所列文献，整理出一份包含 59 款视频注释工具的详尽清单。我们找到了这些工具的网站、下载链接或共享的开源代码；至少也找到了作者的演讲视频或教程视频。我们整理了大多数软件支持的典型功能，以及各工具独有的功能和技术。在这项广泛的调查中，我们提供了筛选出的五款工具的对比表。我们在第6.1节详细说明了筛选方法。我们制作了两种类型的表格：

- 基于表1所示的高层分类体系，以及
- 基于表2所示的功能

作为对比，图3展示了使用3种不同的视频标注工具创建新标注所需步骤的示例。我们在第6.2节的评估部分进行了更全面的逐步比较。

- 左上角是 ANVIL。要选定区域，需双击起始点，再单击端点时间标记。然后右键单击阴影区域，在上下文菜单中选择相应选项，以创建和编辑标签。
- 右上角是 VIAN。在 VIAN 中，右键单击轨道与标尺之间的空白处，然后选择“创建标注”，再选择“文本”作为文本标注的类型。
- 在 VIA 中，先为特定标注创建轨道并为其命名（本例中为“鸟”），然后播放视频直到影片的起始点，并按字母“A”创建固定大小的标注。之后可以返回并调整该标注。

我们通过使用多款事件注释软件来标注视频，或观看作者发布的视频，对这些软件的布局和功能进行了调查。我们注意到以下几点：

SN	功能	描述	Advene	ANVIL	ELAN	VIA	VIAN	FEVA（我们）
0	最后更新	年月	2020年6月	2019年3月	2020年3月	2020年7月	2020年5月	2020年9月
1	软件平台	云端与边缘端	Edge	Edge	Edge	Edge	Edge	云端或边缘端
		原生与网络基于	原生	原生	原生	基于网页	原生	基于网页
		模块化 vs 静态	静态	静态	静态	静态	静态	模块化
2	许可证	Open 源代码与专有相对	开源	专有	开源	开源	开源	开源
		商业版与开放获取	开放获取	开放获取	开放获取	开放获取	开放获取	开放获取
		维护 vs 已过时	维护	维护	维护	维护	维护	维护
3	Cost	Free，低成本，对比昂贵	Free	Free	Free	Free	免费/低成本	Free
		一次性购买与订阅	N/A	N/A	N/A	N/A	N/A	N/A
4	协作	单用户	✓	✓	✓	✓	✓	✓
		多用户（Si-（同时）	✗	✗	✗	✗	✗	✗
		人群	✗	✗	✗	✗	✗	✗
5	目标用户	技术 vs 非技术	技术	技术	技术	技术	技术	Both
		学术 vs 商业	学术	学术	学术	学术	学术	学术
6	输入类型	图像	✗	✗	✗	✓	✗	✗
		视频	✓	✓	✓	✓	✓	✓
		音频	✗	✗	✓	✓	✗	✓
7	标注 Type	对象	✗	✗	✗	✓	✗	✓
		动作	✓	✓	✓	✓	✓	✓
		事件	✓	✓	✓	✓	✓	✓
		混合	✗	✗	✗	✓	✗	✓
8	标注方-法	人工	✓	✓	✓	✓	✓	✓
		自动	✗	✗	✗	✗	✗	✗
		混合	✗	✗	✗	✗	✗	✗
9	标注格式	JSON	✗	✓	✗	✓	✗	✓
		XML	✓	✗	✓	✗	✗	✗
		SQL	✗	✗	✗	✗	✗	✗
		其他?	✓	✓	✓	✓	✓	✗

表 1：视频注释工具分类体系对比表。

SN	功能	描述	Advene	ANVIL	ELAN	VIA	VIAN	FEVA
1	标注类型	对象边界框	✗	✗	✗	✓	✗	✓
		对象掩码	✗	✗	✗	✓	✗	✗
		对象点	✗	✗	✗	✓	✗	✓
		时间事件	✓	✓	✓	✓	✓	✓
2	播放控制	播放 暂停 快进 快退	✓	✓	✓	✓	✓	✓
		速度 +/-	✓	✓	✓	✓	✓	✓
		时间轴跳转	✓	✓	✓	✓	✓	✓
3	预览	缩略图预览	✓	✗	✗	✗	✗	✓
4	标签	多轨道	✓	✓	✓	✓	✓	✓
		轨道分组	✗	✗	✗	✗	✗	✗
		用户定义的标签类型	✓	✓	✓	✓	✓	✓
		显示/隐藏/折叠/展开	✗	✗	✗	✗	✗	✗
5	速度标注	Sudo-Pedal	✗	✗	✗	✓	✗	✓
		转录踏板支持	✗	✗	✗	✗	✗	✓
6	调整大小		✓	✓	✓	✓	✓	✓
7	Move		✓	✓	✓	✓	✓	✓
8	Add		✓	✓	✓	✓	✓	✓
9	删除		✓	✓	✓	✓	✓	✓
10	Edit		✓	✓	✓	✓	✓	✓
11	导入		✓	✓	✓	✗	✓	✓
12	导入其他格式		✓	✓	✓	✗	✓	✓
13	视频支持	MP4	✓	✓	✓	✓	✓	✓
		其他	✓	✓	✓	✓	✓	✗
14	摄像机	多摄像机	✓	✗	✗	✗	✗	✓
		切换视图	✓	✗	✗	✗	✗	✓
		即时切换视图	✓	✗	✗	✗	✗	✓
15	历史记录	撤销/重做	✓	✗	✓	✗	✓	✓
16	搜索	关键词	✓	✗	✓	✗	✓	✓
		按标签类型筛选	✓	✗	✓	✗	✓	✓
17	用户配置	记住/保存	✗	✗	✓	✗	✗	✓
18	模块化/API	插件支持	✓	✗	✓	✗	✓	✗
		全面支持开源	✓	✗	✓	✓	✗	✓
		自定义图层	✗	✗	✗	✓	✓	✗
		自定义轨道	✓	✓	✓	✓	✓	✓
19	图层	显示/隐藏图层	✗	✗	✗	✗	✓	✓
		人体关节关键点支持	✗	✗	✗	✗	✗	✓
		人体边界框支持	✗	✗	✗	✓	✗	✓
		人体掩膜支持	✗	✗	✗	✓	✗	✗
20	导出支持	视频片段	✓	✗	✓	✗	✗	✓
		图像帧	✗	✗	✓	✗	✗	✓
		隐藏式字幕	✓	✗	✓	✗	✗	✓

表2：视频注释工具功能比较表。

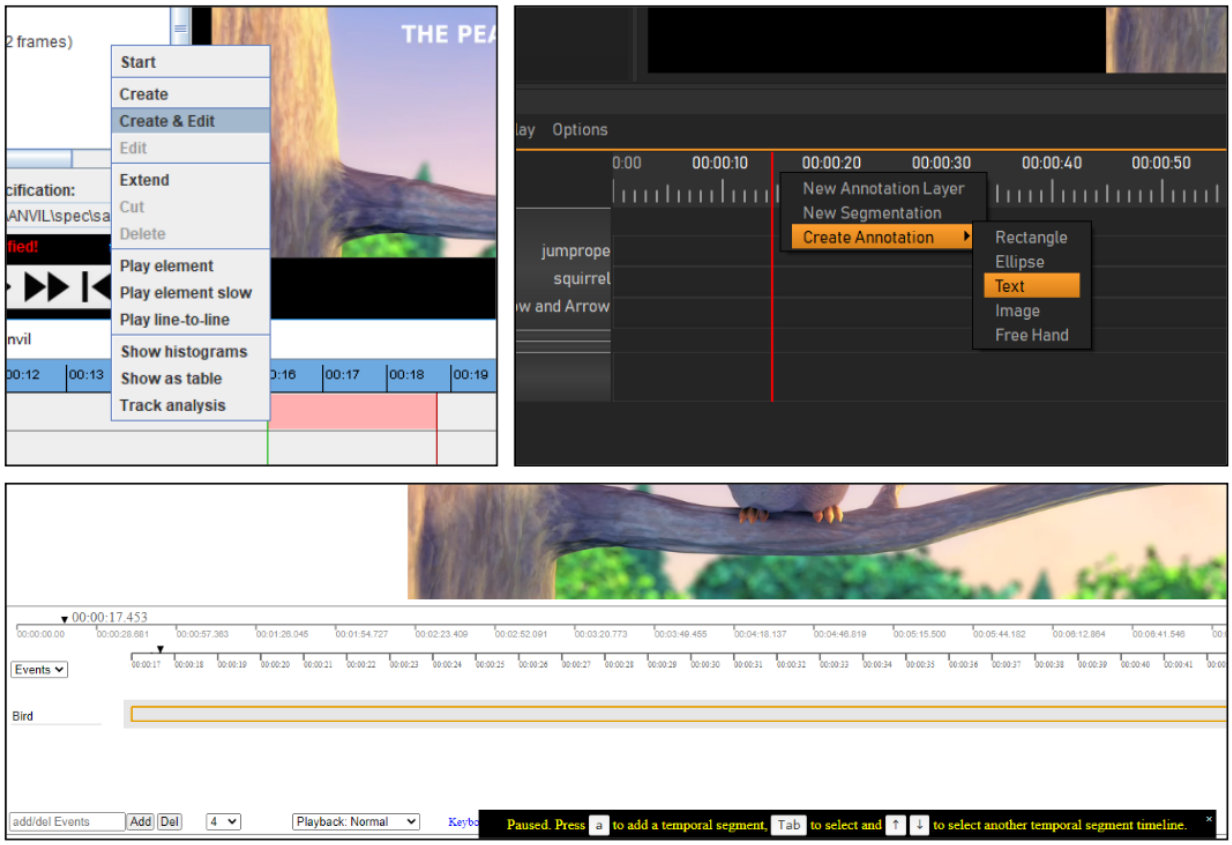


图3：在3种不同的视频标注工具中进行标注所需步骤示例。左上方的截图是 ANVIL，右上方的是 VIAN，底部的截图是 VIA。

• 除 VIA 外，大多数视频标注工具的 [12] 屏幕都被按钮、菜单和窗口占满。虽然重要功能藏在多层菜单项中，但屏幕空间仍未得到充分利用。新的视频标注工具需要简单易懂且易于导航。

• 通过观看视频说明和进行试点测试，我们发现只需要有限的功能和控件，主要包括：a) 项目和数据管理，b) 标注管理，c) 视频导航，d) 标注空间，以及 e) 工具配置。新的视频标注工具需要采用合适的布局，让这些功能一目了然，并去除不必要的功能，或将其安排在不常用的区域。

• 标注可视化功能有限。每个轨道只能用于一个标签 [13], ，而且无法设置重叠时间标签 [8, 10]。只有一些 [12, 9] 允许设置重叠时间标签，但标签难以区分和操作。新的视频标注工具需要自动整理标注，并更有效地利用标注空间。

• 虽然有些工具 [12, 11] 提供了良好的标注用户体验控件，可用于创建和编辑时间位置，但在标注步骤之间必须交替使用鼠标和键盘。有些工具 [8, 9, 10]移动或调整大小标注的方式非常繁琐。有些工具提供了历史记录选项，可用于撤销/重做[9, 10], ，而另一些工具则无法回退用户的错误操作或进行尝试性操作。新的视频注释工具需要提供简单的鼠标控件和默认键盘快捷键，同时允许用户重新定义这些快捷键。

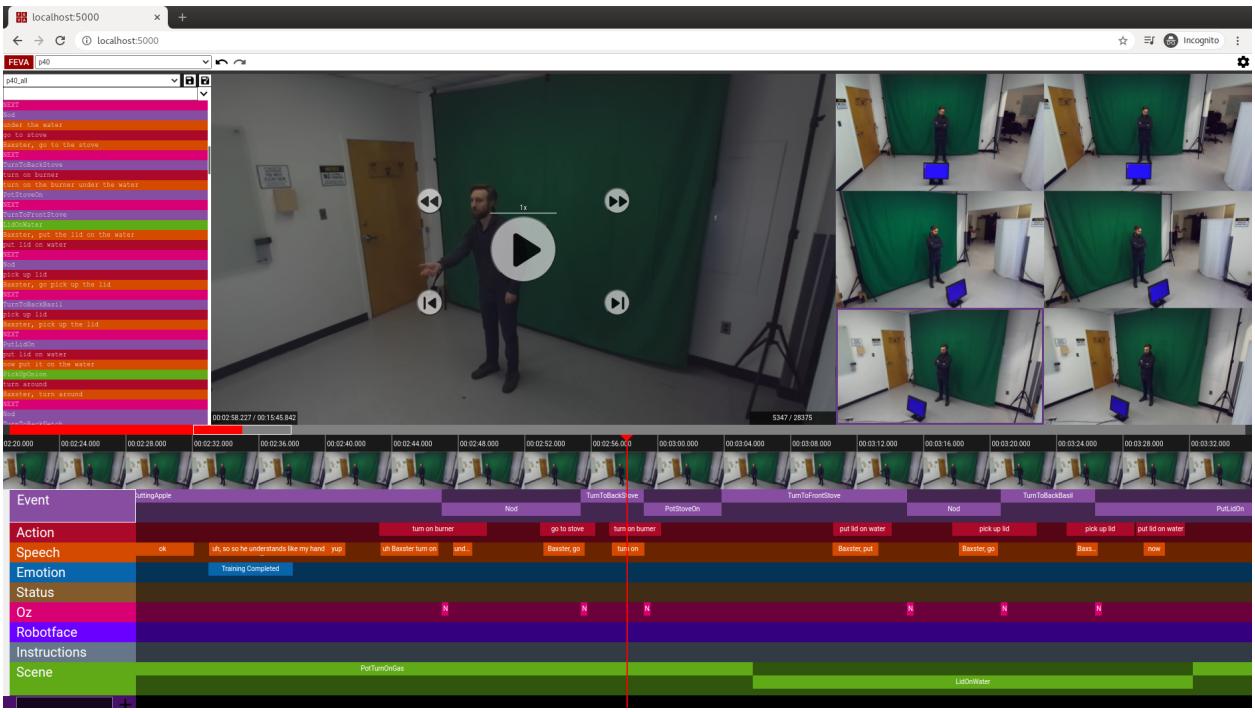
• 除了 VIA [12], 之外，所有工具都要求用户停止播放并进行标注；VIA 则可在视频播放期间提供实时标注。然而，VIA 只允许使用定长标注标记，因此之后还需要调整端点。此外，VIA 的可视化效果不佳，也缺少控制重叠标签的功能。新的视频注释工具需要提供一种快速的实时方式，让用户能够继续标注而无需停止播放。

4 设计考量

根据访谈、文献调查以及对这些工具的使用情况，我们提出了五项有助于开展视频标注活动的设计标准：

- D1. 根据操作工作流程组织空间：功能应按照工作流程的逻辑顺序排列，同时不要偏离常见的软件惯例太远。在当前情境下不需要的功能不应显示或启用，以腾出宝贵的屏幕空间。高频控件和功能应位于屏幕中央。
- D2. 简化高频操作：对于创建标注、微调标注等重复频率很高的操作，应加以优化，使其更容易、更快速地完成。尽可能让这些操作只需按一次键即可完成。尽量只使用一种设备（即不要让某些步骤必须使用键盘，而另一些步骤必须通过鼠标移动来完成，以免在设备间切换时浪费宝贵时间）。在可能的情况下，运用 D1 的原则，尽量减少完成任务所需的用户交互。
- D3. 尽可能利用算法支持：只要可行，就尽量减轻用户的负担，让算法承担繁重工作。例如，在处理时间点时，应考虑用户反应时间，并根据用户输入相对于预期时间的延迟进行调整。此外，允许使用动作检测、人体检测、语音检测和识别等外部模块，可以免去用户手动查找和标注事件的工作。不过，我们应谨记，无论机器学习模块多么出色，算法都不可能达到百分之百的准确率。因此，这些模块只能作为辅助，不能完全依赖它们来生成真实标注。
- D4. 采用行之有效的方法，并改进不奏效的部分： 与其重新发明轮子，不如借鉴其他工具中经用户测试或试点测试证明有效的做法，而不是凭直觉和个人偏好行事。
- D5. 提供灵活性：虽然具备 D4 中经过测试的默认输入方式很不错，但通过鼠标上下文菜单和键盘快捷键提供冗余操作，增加灵活性，可能会让一些用户觉得更加直观。此外，还应允许用户重新定义按键映射，因为每个人都有自己的偏好。

5 FEVA



Fig图 4：FEVA 的屏幕截图，其中正在标注一位与机器人互动的参与者。

这些设计考量塑造了 FEVA 的极简设计，大部分屏幕空间都用于显示视频和标注。受点击器/秒表方法和转录脚踏板的启发，我们创建了速度标注功能。该功能允许用户在播放视频时创建标签：用户可以按下按键标记起点和终点，同时继续观看视频，从而创建多个符合所需时长的标签。系统还会考虑用户的反应时间，并调整标记时间，如图1所示。
速度标注功能支持用户实时进行标注，微调控件则允许用户进行帧级调整，而标签整理器会妥善排列标签，避免彼此重叠，方便直接访问。扁平化用户界面响应迅速，且美观悦目。采用 F形设计 [50], , 初始工作流程项目位于左上方，最常查看的项目位于屏幕中间，互动最多的组件则位于屏幕底部。将最常用的元素放在热点区域，而将摄像头选择和配置按钮等较少使用的元素放在右侧较不显眼的位置，这样除非需要，否则它们不会分散用户注意力。该用户界面提供冗余的用户输入方式，以实现灵活、快速的互动，满足新手和专家用户的需求。底层的上下文相关及配置依赖型用户界面模型只会在需要时呈现。专家用户可以通过媒体速度控制和速度标注器，在不触碰鼠标的情况下以快于实时的速度标注视频。研究人员还可以使用缩放功能，以微观或宏观视角查看和分析标签。FEVA 用户界面能在单个屏幕上显示数量最多且含义清晰的标签。

5.1 工作流程

要创建项目，只需选择视频，FEVA 就会将其导入并创建一个空的标注文件，也称为标签或数据集文件（D1、D2）。系统会加载默认轨道，用户可以在配置中自定义这些轨道（D3）。用户可以添加或移除轨道。要播放或暂停影片，用户可以使用主视频上的媒体播放器叠加层，或按下键盘快捷键“空格键”。用户可以使用鼠标滚轮在时间轴中浏览视频。也可以按箭头键（可同时按住或不按住“ctrl”键），以不同幅度移动视频。用户可以点击并拖动胶片条，在全局时间轴上悬停后点击所需时间，或双击胶片条或局部时间轴窗口（D1、D2、D5）。用户还可以更改影片播放速度。使用全局时间轴白框旁的 + 和 - 图标，可以按不同时间间隔放大或缩小；也可以将鼠标悬停在时间轴上，同时按住 ctrl 并滚动滚轮。

要进行标注，用户可以右键点击轨道，并选择想要创建的标签类型。用户也可以按下键盘上的 ’A’ 键来标记开始位置，再次按下以标记结束位置。参见图1。这称为速度标注，因为无需停止视频即可持续进行标注。速度标注需要进行两遍处理，但在我们的试点测试中，速度标注至少比传统方法快1.5倍。

遵循从左到右、从上到下的布局惯例，加载项目和数据集，以及操作视频和标签等工作流程按该顺序组织。主视频位于屏幕中央，占据最大空间。视频下方是标注轨道，其布局遵循用户注视和动作区域的惯例，以及眼睛和手部的布局。参见图5和图4。

5.2 视图/布局

如图5所示，’b’显示项目选择器，用户可以通过它创建新项目并导入视频。’d’显示标签文件选择器，用于创建、加载、保存、合并、导入和导出标签。用户可以使用 ’e’ 进行搜索，并按类型筛选标签。双击 ’2’ 标签列表中的标签，即可在时间轴 ’j’ 中找到对应的标签。用户可以在 ’h’ 中查看缩略图预览，以及 ’g’ 中的当前时间窗口和 ’f’ 中的全局时间轴。

5.3 控制：用户输入系统

每个可通过鼠标控制的媒体控制功能，都有相应的键盘快捷键。为了兼顾灵活性和效率，图形用户界面面向新手用户，遵循识别而非回忆的原则 [51], , 同时为专家用户提供快捷键，以便更快地进行控制。用户可以点击视频上叠加的播放按钮，或随时按空格键来播放或暂停视频。你也可以选择加快或减慢播放速度。视频播放期间若要创建标注，可以连续按两次字母“A”，分别标记标签的开始和结束。此操作可在配置中自定义。根据你的反应时间进行调整后，系统会创建一个空白标签。你也可以双击轨道上的空白区域，或右键点击轨道并选择要创建的标签类型。虽然这些快捷键的选择遵循了其他视频标注工具的既有标准，但你可以在用户配置中轻松自定义所有快捷键。

导航时，您可以使用鼠标滚动浏览胶片条，也可以单击并拖动胶片条；还可以在胶片条或全局进度条上双击所需位置。您也可以在标签列表中双击某个标签，直接跳转到该标签。用户可以选择自己偏好的操作方式，因为系统同时提供键盘和鼠标操作，以及可由用户重新自定义的键盘快捷键，避免只能依赖单一方式。

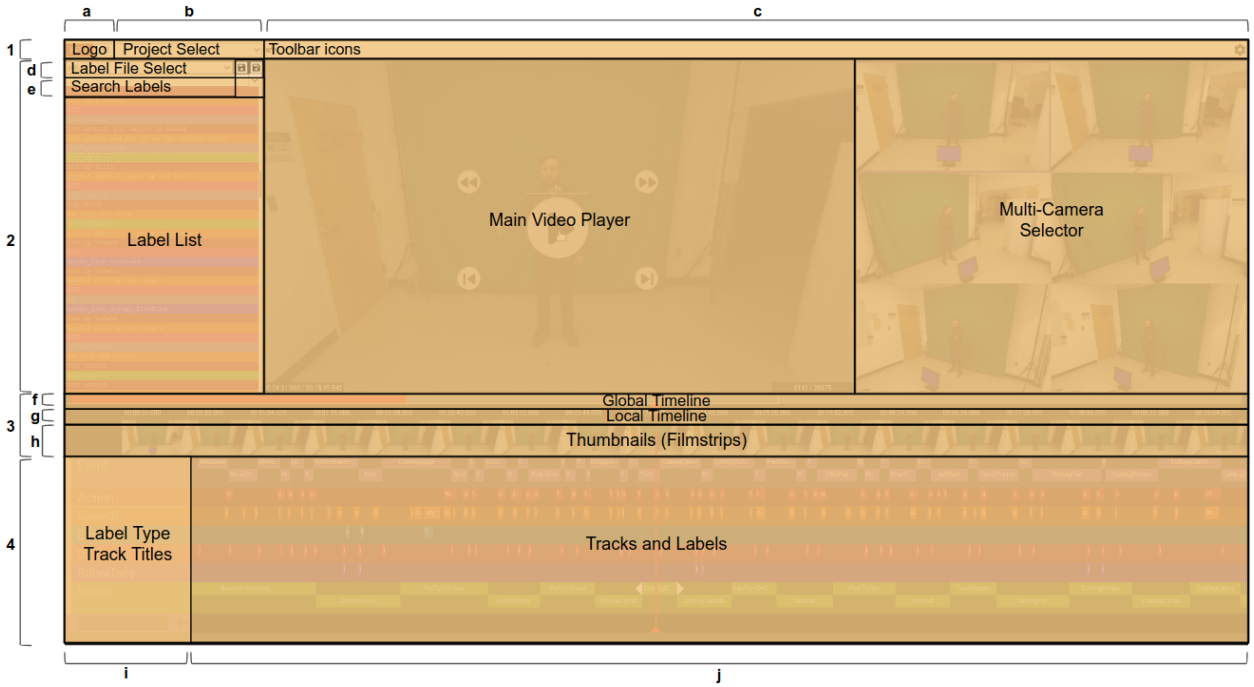


图5：FEVA 屏幕默认布局。黑框区域分别为：1）工具栏；2）标签列表、主视频播放器和多摄像头选择器；3）视频导航时间轴；4）标签轨道。主要组件包括：a）标志；b）项目选择；c）工具栏图标；d）标签数据文件选择；e）搜索栏；f）全局进度条时间轴；g）局部时间轴刻度尺；h）胶片条；i）标签类型；j）轨道和标签。

5.4 模型/状态：底层用户界面支持系统

用户界面通过利用z轴对显示的组件进行分层，并根据上下文改变状态，从而充分利用有限的屏幕空间（x-y平面）。这些上下文包括视频是否正在播放、是否选中了标签、是否正在编辑标签、鼠标是否悬停在某个组件上，或配置中是否启用了某项功能。每个区域都紧凑地整合了各种功能，其设计依据是用户自然会赋予这些组件的可供性，因此使用起来直观。视频播放器区域就是一个例子。视频播放时，用户只能看到视频；鼠标指针悬停时，则会显示媒体控件、当前时间和帧编号，以及图层控制。您还可以在配置中启用或隐藏视频、人体关键点 [52], 、人体或物体的边界框、分割掩码等显示功能，并可由研究人员进一步扩展和自定义。

6 评估

为了评估 FEVA，我们通过两项研究将 FEVA 与现有的最先进（SOTA）视频标注工具进行了比较。

- 互动基准测试：为了评估使用每种视频标注工具时用户能够多快完成标注的理论极限，我们统计了完成各项任务所需的用户输入次数。
- 用户研究：为了根据用户对每种视频注释工具的主观工作负荷评估用户体验，我们开展了用户研究，收集用户根据自身体验提供的反馈。

6.1 最先进视频注释工具选择方法

我们首先整理了一份引用率较高、可供下载和使用的视频注释工具清单。出于实际考虑，我们只纳入了支持时间标注，且无需采取极端措施即可下载、安装并运行的软件。因此，VCode [11], SVAT [53], 和 VACA [35] 未能纳入。由于缺少开始时间和结束时间等功能，ToolScape [47], HistoryTracker [46], 和 CASAM [54] 等过于专用的工具也从清单中移除。由于我们开展的是单用户研究，因此排除了 Glance [55]和 CoAT [42], 等专注于众包的工具。根据这些标准，我们将研究中用于比较的工具范围缩小到

这些工具。EagleView [36] 运行时极不稳定，因此我们无法对其进行测试。最终，我们将范围缩小至 Advene [9], ANVIL [8], ELAN [10], VIA [12],和 VIAN [13]。

6.2 交互基准测试

为了比较使用各类视频标注工具完成特定任务所需的步骤，我们统计了鼠标点击、双击、鼠标移动和键盘按键次数，并计算总和，如表3所示。如果完成某项任务有多种方法，我们则采用该工具中最快的方法。例如，如果可以按 Ctrl+N 创建新项目（按键操作计数为 = 2），也可以将鼠标移至主文件菜单，点击“文件”，再将鼠标移至“新建项目”，然后点击该菜单项（鼠标移动 = 2 次，鼠标点击 = 2 次，总计 = 4 次），那么我们就取两者中较小的数值。

表3列出了评估中考察的15项任务，包括基本设置、标签创建和操作任务。我们选取了大多数视频标注工具都能完成的任务。如果某个视频标注工具缺少某项功能，我们就为其计入与其他视频标注工具比较时得到的最高次数。例如，[12]不支持“撤销”或“重做”，因此计为2次。
如表 3 所示，FEVA 所需的输入次数显著少于当前最先进水平。平均而言，FEVA 所需输入比当前最先进水平少 36%。根据 T 检验，FEVA 所需输入显著少于除 VIA 以外的所有工具；与 VIA 的差异则未达到统计显著性。

SN	任务	Advene	ANVIL	ELAN	VIA	VIAN	FEVA
1	创建项目 + 导入视频	7	8	8	7	14	**3**
2	创建单个标签	4	6	7	**2**	5	**2**
3	创建多个标签	8	12	14	**3**	9	**3**
4	创建并命名标签	**4**	9	7	7	9	7
5	编辑标签	4	7	**3**	**3**	**3**	5
6	调整标签大小	6	5	6	**4**	**4**	**4**
7	移动标签	6	12	6	5	**4**	**4**
8	更改标签类型	6	6	6	6	6	**4**
9	删除标签	**3**	5	4	**3**	**3**	**3**
10	查找标签	3	3	5	5	3	**2**
11	保存标签	**2**	**2**	**2**	**2**	**2**	**2**
12	加载标签	**4**	**4**	**4**	**4**	8	**4**
13	浏览视频	2	6	2	**1**	2	**1**
14	播放/暂停视频	2	**1**	2	**1**	**1**	**1**
	仅播放带标签的视频	6	5	4	3	6	**2**
15	撤销/重做	**2**	2	**2**	2	**2**	**2**
	总分	69	93	82	58	81	**49**
	FEVA 快于	**29%**	**47%**	**40%**	**16%**	**40%**	
	p值	0.0255	0.0013	0.0149	0.1321	0.0223	

表3：使用各软件中最快方法完成的任务列表（适用时使用快捷键）。每个数字表示鼠标点击、双击、移动和按键次数的累计总和。最后一行显示 FEVA 比当前最先进水平快多少（以百分比（%）表示），以及 T检验 p值。

6.3 用户研究

为了评估用户体验，我们开展了一项用户研究，让参与者以轮换方式使用两种视频标注工具：FEVA 和另外一种选定的当前最先进水平视频注释工具（Advene、ANVIL、ELAN、VIA 和 VIAN）。我们通过让下一位参与者使用相反的顺序来平衡这两种工具的使用顺序。由于 COVID-19 相关规定，我们通过 Zoom 的远程共享屏幕和控制功能开展研究。这给用户体验带来了一些延迟。不过，由于两种工具都以远程方式使用，我们认为延迟对结果的影响并无显著差异。我们向参与者提供事件发生的大致时间范围，以及要标注事件的清晰描述。例如，如图6所示：“请标注 4 分 20 秒至 4 分 40 秒之间的兔子跳绳。”参与者完成了约 24 项任务，直到他们放弃使用该工具。在完成所有

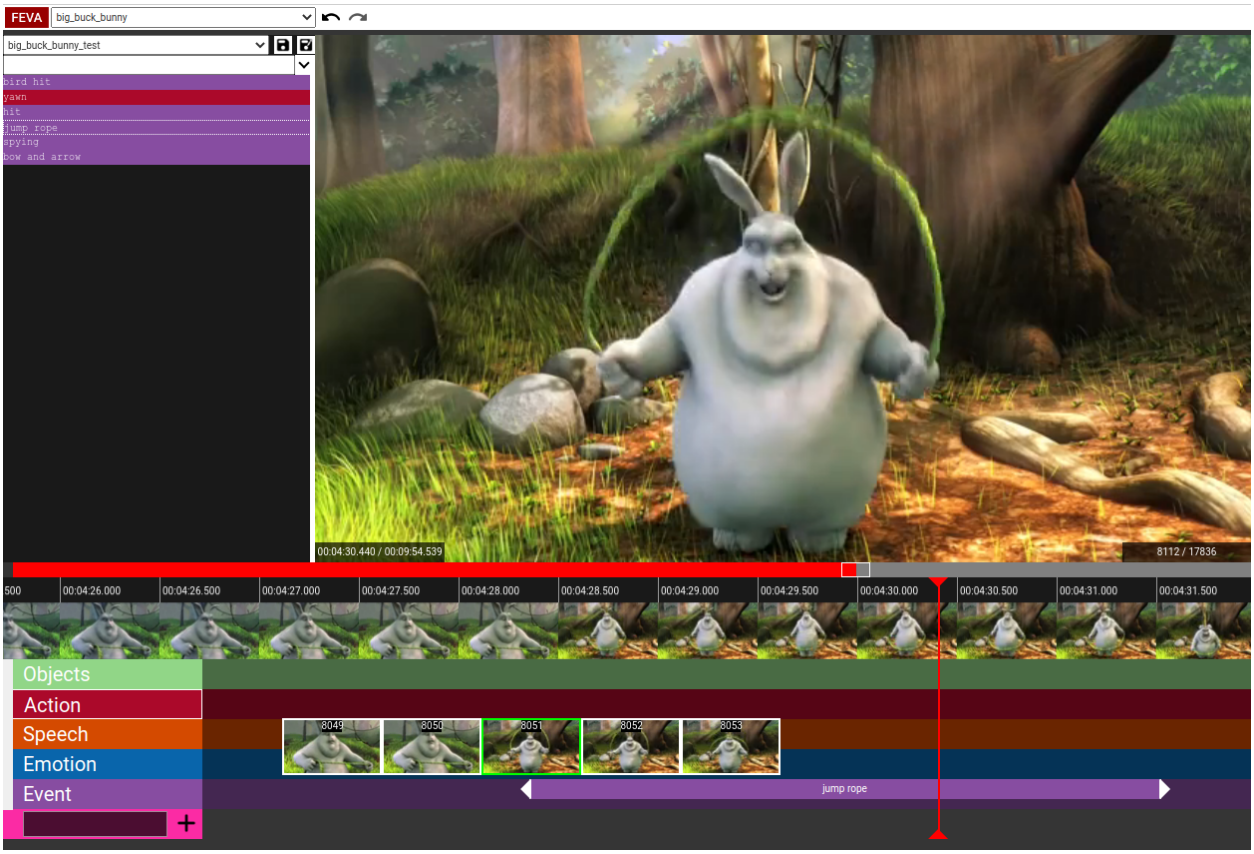


图6：FEVA 中的一项任务示例：参与者被要求标注 04:28 分钟至 04:31 分钟之间兔子跳绳的事件。

在第一款软件中完成任务后，他们填写了 NASA任务负荷指数 [56] 问卷，并使用5点李克特量表评分。随后，他们使用第二款工具重复了这一流程。

6.3.1 参与者

我们通过电子邮件和社交媒体论坛，从大学社区招募了 34 名参与者。在我们针对 N=32 的研究中，参与者中男性占 53%，女性占 47%，平均年龄为 30.4 +/- 5.9 岁；其中 84% 没有视频标注经验，66% 没有视频编辑经验。由于 Zoom 连接问题，两名参与者不得不被剔除，因此不计入 N=32。

6.3.2 流程

在本研究中，我们先让标注人员观看一段简短的培训视频，并进行为期3分钟的培训，学习媒体控件、视频时间线导航以及创建和编辑标注的基础知识；随后，他们针对每个项目进行练习试验，由研究协调员解答疑问。之后，他们又花了2分钟自行探索该工具。接下来的5分钟里，参与者练习研究人员为每人单独安排的任务，并可随时提问。待他们熟悉操作后，我们随机向他们分配以下列表所示的四类任务，每种类型的任务至少重复三次。所选任务是标注人员在视频标注过程中必须完成的最基本且最常见的任务。这些任务的复杂程度各不相同，有些需要组合完成多个基本步骤。例如，有些任务只要求参与者将视频定位到2分20秒处；另一些任务则要求参与者标注特定时间段内连续发生的三个事件，并为其恰当命名。完成任务没有时间限制。他们使用的是标准的免费“Big Buck Bunny”视频，该视频时长约10分钟，分辨率为720p。图6展示了一个任务示例。使用第一个工具完成所有任务后，参与者填写NASA TLX工作负荷问卷，随后使用第二个视频注释工具重复完成这些任务。

我们在用户研究中进行了以下四类基本任务：

• 浏览视频 a）播放/暂停视频，b）跳转到时间轴上的特定位置，以及 c）跳转到特定标签所在的精确位置。

• 标签创建 a）在特定时间以特定长度创建新标签，b）当参与者表现出特定行为时创建标签（例如，角色打哈欠、吃苹果等），以及 c）连续创建多个标签。

• 标签内容操作 a）为已创建的标签编写文本标注，以及 b）修改标注文本。

• 标签时间操作 a）将标签移动指定的秒数，以及 b）调整标签大小，以更改其开始时间或结束时间，使其与视频中人物的特定行为相匹配。

6.3.3 结果

在本研究中，与当前最先进水平相比，用户使用 FEVA 时平均感受到的心理需求降低了**46%** ($p < 0.00003$)，体力需求降低了**41%** ($p < 0.00187$)，所需努力程度降低了**34%** ($p < 0.00324$)，挫折感降低了**62%** ($p < 0.00147$)。时间需求和绩效水平指标的差异并不显著。我们认为，这是因为研究期间没有设置时间限制；除了一名使用 VIAN 的用户选择放弃外，其他用户均完成了所有任务。

FEVA	心理需求	体力需求	时间需求	绩效水平	努力程度	挫折感程度
平均值（n=32）	1.8	1.5	1.8	4.3	2.3	1.7

表4：展示了FEVA在采用5点李克特量表的NASA任务负荷指数上的平均得分 e.

FEVA 对比（%）	心理需求	身体需求	时间需求	绩效水平	努力程度	挫折感程度
Advene（n=5）	75%	57%	0%	14%	67%	183%
ANVIL（n=8）	25%	36%	-8%	3%	17%	6%
ELAN（n=8）	29%	42%	13%	12%	38%	83%
VAI（n=8）	44%	36%	6%	11%	22%	36%
VIAN（n=3）	120%	40%	20%	17%	83%	140%
平均值（n=32）	**46%**	**41%**	5%	10%	**34%**	**62%**

表5：展示了 FEVA 与其他视频注释工具的比较结果。正数表示人们认为 FEVA 在 NASA任务负荷指数的六个维度上比其他视频标注工具好多少，负数则表示差多少。

6.4 用户反馈

6.4.1 优点

平均而言，用户有 88% 的时间认为 FEVA 更直观，有 91% 的时间认为 FEVA 更易于使用。用户最喜欢的功能包括速度标注、微调、“更凉爽的感觉”、定位标签和标签回放。少数用户希望在使用第二种工具后，能够返回并修改自己对第一种工具的反馈。这种情况通常发生在他们使用 FEVA 后，觉得自己给第一种工具打分过高时。相反，如果使用顺序颠倒，则不会出现这种情况。一位用户说，他想成立一个粉丝俱乐部，还想自愿参与标注，因为这“太有趣了”。

6.4.2 不良之处

一位用户提到，他们更喜欢传统的窗口式用户界面来处理严肃工作，因此觉得 FEVA 看起来太像主流的平板应用了。另一位用户表示：“虽然我觉得它对我来说没问题，但我不认为我妈妈能使用这两种工具中的任何一种。所以我给两者都打了低分。”一些用户抱怨说，编辑标签后还得按回车键确认，令他们不喜欢。点击其他地方会导致刚输入的内容丢失，这种取消功能并不受欢迎。

6.4.3 不足之处

不过，大多数困惑都与全局时间轴和局部时间线有关，因为两者需要明确区分，却共用了同一个预览组件。参与者通过 Zoom 远程操控研究协调员电脑上的视频标注工具用户界面时，发现操作效果存在延迟。一些用户抱怨说，Zoom 延迟导致用户界面更新不够快。“可能是因为我通过 Zoom 操控你的电脑，但巨大的延迟让我更难调整标签大小。”

7 实现

7.1 框架与依赖项

我们使用了 ReactJS [57], ，这是一款高效的、基于组件的 JavaScript 库，并将架构设计得轻量且响应迅速，因此能够在大多数人的电脑上运行。安装时只需依赖 Python 和 Flask。前端采用标准的 HTML、JavaScript、ReactJS 和 CSS。我们设计所有控件时都兼顾了性能和自定义灵活性。图5和第 5 节详细介绍了用户界面布局。

7.2 架构

FEVA 采用了许多基于网络的应用程序常见的简单服务器—客户端架构。服务器端运行在 Python 3.5x 或更高版本上，并使用 Flask 作为网络服务器。服务器端主要负责在客户端应用程序发出请求时提供数据（网页内容、标注数据和视频流）。对于 FEVA 来说，大多数模块和设计都位于客户端。附录 8 的框图展示了所有模块及其与其他模块互动的更多细节。

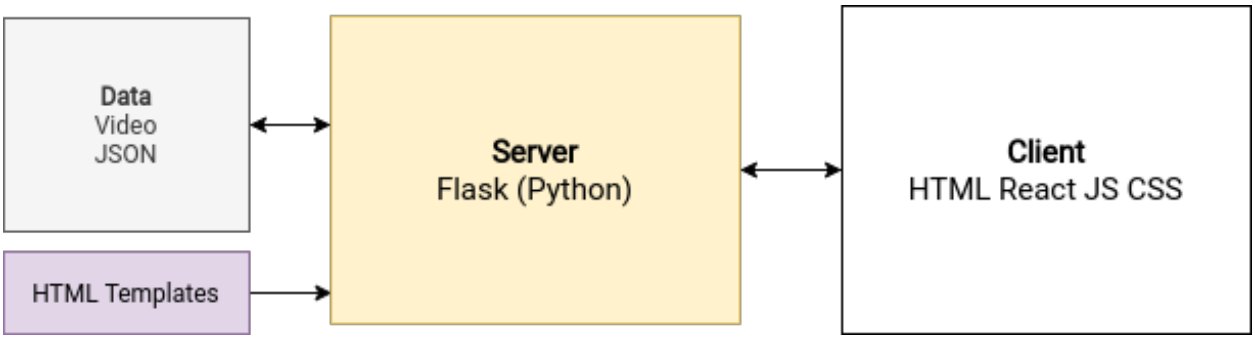


图 7：FEVA 客户端—服务器架构图。附录 C 中提供了客户端不同模块的详细框图。

8 讨论

这是我们工具 FEVA 发布的首个版本，我们着重构建其基础功能，同时优化用户界面和互动，以便更快速、直观、稳健且准确地标注事件。尽管这些早期结果看起来颇有希望，但仍有更多研究问题有待进一步探讨。

8.1 用户研究

在我们的试点研究和用户研究中，我们开展了一项短时研究。在用户研究中，我们假设 15 分钟足以让参与者学习并练习进行标注，因此我们据此设计了首次比较多个视频标注工具的评估。不过，我们还需要通过开展纵向用户研究来进一步推进研究，以了解随着用户越来越熟悉该软件，我们的设计会对用户产生怎样的影响。

8.2 用户输入

在输入领域，我们现阶段将键盘和鼠标/触控板视为交互设备。不过，我们还需要将范围扩展到其他类型的输入，例如触控、言语和手势，以探索多模态方法的潜在优势。

8.3 目标用户

随着越来越多的公众参与编码过程，我们将本研究的重点放在让任何人都能参与编码过程上。未来开展研究以收集经验丰富的编码人员的反馈，将有助于了解他们如何使用 VAT。

8.4 VAT 基准测试研究

在第 3 节的研究中，我们统计了输入次数，因为试点研究显示，点击次数与完成任务所需时间之间存在相关性。未来应开展更全面的研究，将鼠标移动和完成任务所需时间纳入考量，以反映用户困惑程度，并更准确地衡量任务完成情况。

8.5 隐式

在第 6.3 节中，我们着重关注用户的感知，以反映他们的体验。未来的研究应纳入更多用于评估用户困惑程度、表现和成功情况的隐式定量测量。我们还希望开展更多研究，以了解标签的一致性。

8.6 布局

根据用户反馈、从观察中吸取的经验，以及与现有工具进行比较的过程，我们发现还有不少改进空间。多摄像头选择器布局占用了大量屏幕空间。许多用户觉得全局时间轴与局部时间线和缩略图视图之间没有任何分隔，布局过于紧凑，令人困惑，也更难适应。我们计划在后续版本中重新设计这些体验。在下一次发布中，我们将着重把握几个优化机会，让 FEVA 运行得更快。

8.7 未来工作

Bey除了渐进式改进外，我们还计划在未来推出一些关键功能 e:

- 自适应：所有输入控件都是线性的。我们计划探索动态且自适应的控制系统，以采用各种新的交互技术，从而加快标注速度并提供更直观的体验。
- 可扩展：为开源社区提供更便捷的工作流程，以扩展功能。
- AI辅助：虽然目前获得了一些算法支持，但仍需更好地与机器学习和深度神经网络模型集成。我们将进一步研究如何利用 AI 增强标注过程，同时探索如何让用户了解这些模型的不准确性、不确定性和继承偏见。
- 远程视频：虽然我们的内部原型支持 YouTube，但在将其无缝用作本地 MPEG视频的替代方案之前，我们还需要探索优化机会。我们也将探索其他在线流媒体平台。
- 案例研究：我们正在与一些研究实验室合作，评估 FEVA 在视频标注中的表现[58], 我们希望邀请其他感兴趣的研究实验室试用 FEVA，与我们合作，并共同发展壮大，满足迄今研究尚未发现的需求。

9 结论

我们推出了一款全新的事件视频标注工具，采用简化的交互技术和动态用户界面，并根据工作流程和使用频率，对上下文中可见且处于激活状态的功能进行组织。借助快速标注等功能，用户可以实时准确地标注视频。简化的入门引导和工作流程，让研究人员只需很少时间便可完成设置并开始标注视频。我们在 GitHub 上发布了 FEVA 的源代码，供所有人试用并进一步扩展其功能。社区成员还可以在 GitHub 页面上找到项目示例、教程视频、用于获取支持的 GitHub问题以及未来更新。随着我们不断扩展案例研究，我们欢迎更多研究人员使用 FEVA；如果您有意合作，也欢迎联系我们。

致谢

我们感谢 Chethan Parameshwara、Levi Burner、Lindsay Little，以及来自马里兰大学和感知与机器人组的同仁提供宝贵的反馈并参与讨论。我们特别感谢所有项目贡献者：Johnny Chiu、Rachelle Sims、John Gao、Leya Abraham、Vikram Sehgal、Swagata Chakroborty、Lucas Stuart 和 Lin Chen。我们衷心感谢美国国家科学基金会通过 OISE 2020624 项目提供的资助。

参考文献

[1] Pat Broadhead. 从幼儿园到一年级的合作游戏与学习。早期儿童的游戏与学习，第43–60页，2010年。[2] Grace Megumi Chen、Keith Jonathon Yoder、Barbara Lynn Ganzel、Matthew S Goodwin 和 Matthew Kenneth Belmonte。利用重复行为吸引注意力并促进学习：一种自闭症新疗法的探索性分析。《心理学前沿》，3:12，2012年。[3] Virginia K Choi、Snehesh Shrestha、Xinyue Pan 和 Michele J Gelfand。危险来袭时：一种追踪美国集体应对威胁情况的语言学工具。《美国国家科学院院刊》，119(4)，2022年1月。[4] Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Sanja Fidler、Antonino Furnari、Evangelos Kazakos、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 和 Michael Wray。EPIC-KITCHENS 数据集：收集、挑战与基准。*IEEE*模式分析与机器智能汇刊（*TPAMI*），2020年。[5] Andreas Geiger、Philip Lenz、Christoph Stiller 和 Raquel Urtasun。视觉与机器人学相遇：KITTI 数据集。《国际机器人研究杂志》（*IJRR*），2013年。[6] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。Audio Set：一种用于音频事件的本体与人工标注数据集。载于*2017*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第776–780页。IEEE，2017年。[7] Sami Abu-El-Haija、Nisarg Kothari、Joonseok Lee、Paul Natsev、George Toderici、Balakrishnan Varadarajan 和 Sudheendra Vijayanarasimhan。YouTube-8M：大规模视频分类基准。*arXiv*预印本 *arXiv:1609.08675*，2016年。[8] Michael Kipp。ANVIL：视频标注研究工具。语料库音系学手册，第420–436页，2014年。[9] Olivier Aubert、Yannick Prié 和 Daniel Schmitt。Advene：一种可定制的超视频创作工具：案例研究。载于*2012*年*ACM*文档工程研讨会论文集，第79–82页，2012年。[10] Peter Wittenburg、Hennie Brugman、Albert Russel、Alex Klassmann 和 Han Sloetjes。ELAN：多模态研究的专业框架。载于第五届国际语言资源与评测会议（*LREC 2006*），第1556–1559页，2006年。[11] Joey Hagedorn、Joshua Hailpern 和 Karrie G Karahalios。VCode和VData：展示支持视频标注工作流程的新框架。载于高级可视化界面工作会议论文集，第317–321页，2008年。[12] Abhishek Dutta 和 Andrew Zisserman。用于图像、音频和视频的VIA标注软件。载于第*27*届*ACM*国际多媒体会议论文集，第2276–2279页，2019年。[13] Gaudenz Halter、Rafael Ballester-Ripoll、Barbara Flueckiger 和 Renato Pajarola。VIAN：用于电影分析的可视化标注工具。载于《计算机图形学论坛》，第38卷，第119–129页。Wiley 在线图书馆，2019年。[14] Charlotte Ward、Kevin Baker、Sarah Marks、Dawit Getachew、Tedila Habte、Cindy McWhorter、Paul Labarre、Jonathan Howard-Brand、Nathan P Miller、Hayalnesh Tarekegn 等。确定自动呼吸频率计数器与检测儿童肺炎症状参考标准之间的一致性：埃塞俄比亚横断面研究方案。《*JMIR* 研究方案》，9(4):e16531，2020年。[15] Rob Tannen。专题：研究人员与工具不匹配——改善用户研究人员与技术之间的契合度。《*Interactions*》，15(5):74–78，2008年。[16] Michaela Regneri、Marcus Rohrbach、Dominikus Wetzel、Stefan Thater、Bernt Schiele 和 Manfred Pinkal。将动作描述与视频关联。《计算语言学协会汇刊》，1:25–36，2013年。[17] Carl Vondrick、Donald Patterson 和 Deva Ramanan。高效扩展众包视频标注。《国际计算机视觉杂志》，101(1):184–204，2013年。

[18] Hugo Jair Escalante、Víctor Ponce-López、Jun Wan、Michael A Riegler、Baiyu Chen、Albert Clapés、Sergio Escalera、Isabelle Guyon、Xavier Baró、Pål Halvorsen 等。Chalearn多媒体挑战联合竞赛：超越视觉分析的综述。载于*2016*年第*23*届国际模式识别会议（*ICPR*），第67–73页。IEEE，2016年。[19] Michael Riegler、Martha Larson、Concetto Spampinato、Pål Halvorsen、Mathias Lux、Jonas Markussen、Konstantin Pogorelov、Carsten Griwodz 和 Håkon Stensland。Right inflight?：用于探索自动预测适合特定观看情境的电影的数据集。载于第七届国际多媒体系统会议论文集，第1–6页，2016年。[20] Walter S Lasecki、Mitchell Gordon、Winnie Leung、Ellen Lim、Jeffrey P Bigham 和 Steven P Dow。探索众包行为视频编码中的隐私与准确性权衡。载于第*33*届*ACM*计算机系统人因年度会议论文集，第1945–1954页，2015年。[21] Sungjoo Park 和 Chang Mo Yang。用于生成真实标注信息的交互式视频标注工具。载于*2019*年第十一届国际泛在与未来网络会议（*ICUFN*），第552–554页。IEEE，2019年。[22] Carl Vondrick、Donald Patterson 和 Deva Ramanan。高效扩展众包视频标注。《国际计算机视觉杂志》，101(1):184–204，2013年。[23] Simone Bianco、Gianluigi Ciocca、Paolo Napoletano 和 Raimondo Schettini。用于人工、半自动和自动视频标注的交互式工具。《计算机视觉与图像理解》，131:88–99，2015年。[24] VoTT：视觉对象标记工具：一款用于从图像和视频构建端到端目标检测模型的Electron应用，2019年10月。[25] Tewodros A Biresaw、Tahir Nawaz、James Ferryman 和 Anthony I Dell。ViTBAT：视频跟踪与行为标注工具。载于*2016*年第*13*届*IEEE*高级视频与信号监控国际会议（*AVSS*），第295–301页。IEEE，2016年。[26] Ajay Mishra、Yiannis Aloimonos 和 Cornelia Fermüller。面向机器人学的主动分割。载于*2009*年*IEEE/RSJ*智能机器人与系统国际会议，第3133–3139页。IEEE，2009年。[27] Ching L Teo、Cornelia Fermüller 和 Yiannis Aloimonos。二维曲线镜像对称结构的检测与分割。载于*IEEE*计算机视觉国际会议论文集，第1644–1652页，2015年。[28] Francisco Barranco、Cornelia Fermüller 和 Eduardo Ros。使用基于事件的传感器进行实时聚类和多目标跟踪。载于*2018*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第5764–5769页。IEEE，2018年。[29] Bryan C Russell、Antonio Torralba、Kevin P Murphy 和 William T Freeman。LabelMe：用于图像标注的数据库和基于网页的工具。《国际计算机视觉杂志》，77(1-3):157–173，2008年。[30] Cornelia Fermüller 和 Yiannis Aloimonos。跟踪有助于三维运动估计。《生物控制论》，67(3):259–268，1992年。[31] Anton Mitrokhin、Cornelia Fermüller、Chethan Parameshwara 和 Yiannis Aloimonos。基于事件的运动目标检测与跟踪。载于*2018*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第1–9页。IEEE，2018年。[32] Cornelia Fermüller、Fang Wang、Yezhou Yang、Konstantinos Zampogiannis、Yi Zhang、Francisco Barranco 和 Michael Pfeiffer。操作动作预测。《国际计算机视觉杂志》，126(2):358–374，2018年。[33] Eadom Dessalene、Chinmaya Devaraj、Michael Maynord、Cornelia Fermüller 和 Yiannis Aloimonos。基于第一人称视频中的接触表征预测动作。*IEEE*模式分析与机器智能汇刊，2021年。[34] Stamatia Dasiopoulou、Eirini Giannakidou、Georgios Litos、Polyxeni Malasioti 和 Yiannis Kompatsiaris。语义图像与视频标注工具综述。载于知识驱动的多媒体信息提取与本体演化，第196–239页。Springer，2011年。[35] Brandon Burr。VACA：定性视频分析工具。载于*CHI'06*计算机系统人因扩展摘要，第622–627页，2006年。[36] Frederik Brudy、Suppachai Suwanwatcharachat、Wenyu Zhang、Steven Houben 和 Nicolai Marquardt。EagleView：用于可视化和查询人与设备空间交互的视频分析工具。载于*2018*年*ACM*交互表面与空间国际会议论文集，第61–72页，2018年。

[37] Matthew Evanusa、Snehesh Shrestha、Vaishnavi Patil、Cornelia Fermüller、Michelle Girvan 和 Yiannis Aloimonos。面向真实世界时序数据的深读出随机循环神经网络。《SN计算机科学》，3(3)，2022年5月。[38] Yezhou Yang、Cornelia Fermüller、Anupam Guha 和 John Aloimonos。用于理解人类操作动作的认知系统。载于《第二届认知系统进展年会》（ACS），第2卷。Citeseer，2013年。[39] Yezhou Yang、Yiannis Aloimonos、Cornelia Fermüller 和 Eren Erdal Aksoy。学习操作动作的语义。arXiv预印本 arXiv:1512.01525，2015年。[40] Ching L Teo、Yezhou Yang、Hal Daumé、Cornelia Fermüller 和 Yiannis Aloimonos。迈向能够视觉感知的 Watson：面向机器人的语言引导动作识别。载于《2012年IEEE国际机器人与自动化会议》，第374–381页。IEEE，2012年。[41] Konstantinos Zampogiannis、Yezhou Yang、Cornelia Fermüller 和 Yiannis Aloimonos。通过介词落地学习操作动作的空间语义。载于《2015年IEEE国际机器人与自动化会议》（ICRA），第1389–1396页。IEEE，2015年。[42] Aziret Satybaldiev、Peter Hevesi、Marco Hirsch、Vitor Fortes Rey 和 Paul Lukowicz。CoAT：一种基于网页的协作标注工具。载于《2019年ACM普适与泛在计算国际联合会议》附属会议论文集及《2019年ACM可穿戴计算国际研讨会》论文集，第814–818页，2019年。[43] Dalton Bennett、Sarah Cahlan、Aaron C. Davis 和 Joyce Sohyun Lee。特朗普拍照秀前的镇压行动，2020年（访问于2020年9月15日）。[44] Bradford Hosack。VideoAnt：将在线视频标注扩展到内容交付之外。《TechTrends》，54(3)：45–49，2010年。[45] Fabio Sulser、Ivan Giangreco 和 Heiko Schuldt。面向体育视频的基于众包的语义事件检测与视频标注。载于《2014年ACM多媒体众包国际研讨会论文集》，第63–68页，2014年。[46] Jorge Piazentin Ono、Arvi Gjoka、Justin Salamon、Carlos Dietrich 和 Claudio T Silva。HistoryTracker：最大限度减少棒球比赛标注中的人机交互。载于《2019年CHI计算机系统人因会议论文集》，第1–12页，2019年。[47] Juho Kim。ToolScape：提升教程视频的学习体验。载于《CHI '13计算机系统人因扩展摘要集》，CHI EA '13，第2707–2712页，美国纽约州纽约市，2013年4月。计算机协会。[48] Daniel Levin。Timelinely，2020年（访问于2020年9月15日）。[49] Sylvain Lagrue、Nathalie Chetcuti-Sperandio、Fabien Delorme、Chau Ma Thi、Duyen Ngo Thi、Karim Tabia 和 Salem Benferhat。一种基于本体网络应用的非物质文化遗产舞蹈视频标注工具。载于《第一届多媒体遗产内容结构化与理解研讨会论文集》，第75–81页，2019年。[50] Jakob Nielsen。网页内容阅读中的F形模式，2006年4月。[51] Jakob Nielsen。用户界面设计的10条可用性启发式原则，1994年4月。[52] Z. Cao、G. Hidalgo Martinez、T. Simon、S. Wei 和 Y. A. Sheikh。OpenPose：利用部件亲和场进行实时多人二维姿态估计。《IEEE模式分析与机器智能汇刊》，2019年。[53] Herwig Rehatschek、Werner Bailer、Helmut Neuschmied、Sandra Ober 和 Horst Bischof。一款支持视频标注与分析的工具。na，2007年。[54] Robert J Hendley、Russell Beale、Chris P Bowers、Christos Georgousopoulos、Charalampos Vassiliou、Petridis Sergios、Ralf Moeller、Eric Karstens 和 Dimitris Spiliotopoulos。CASAM：多媒体的人机协作标注。《多媒体工具与应用》，70(2)：1277–1308，2014年。[55] Walter S Lasecki、Mitchell Gordon、Danai Koutra、Malte F Jung、Steven P Dow 和 Jeffrey P Bigham。Glance：借助人群快速编码行为视频。载于《第27届ACM用户界面软件与技术年会论文集》，第551–562页，2014年。[56] Sandra G Hart。NASA任务负荷指数（NASA-TLX）：20年之后。载于《人因与工效学学会年会论文集》，第50卷，第904–908页。Sage Publications，Sage CA：洛杉矶，加利福尼亚州，2006年。[57] React——用于构建用户界面的JavaScript库，2013年5月。[58] Snehesh Shrestha、Cornelia Fermüller、Tianyu Huang、Pyone Thant Win、Adam Zukerman、Chethan M. Parameshwara 和 Yiannis Aloimonos。AIMusicGuru：音乐辅助的人体姿态纠正。arXiv预印本 arXiv，2022年。

FEVA：快速事件视频标注工具

A 在线资源

请访问FEVA网站 http://www.snehesh.com/feva，了解更多信息、获取代码、教程视频、示例及最新动态。

B 视频标注工具调查列表

- 用于第一人称视频实时手动标注的形成性研究
- 基于 XML词典的多级视频标注工具
- 用于为智能视频监控系统生成真实标注的半自动视频标注工具
- Advene
- AIBU
- 一种用于视频标注的创新型基于网络的协作平台
- 一种基于本体网络应用的非物质文化遗产舞蹈视频标注工具
- ANVIL
- atlas.ti
- 增强工作室：面向物理治疗教育的运动人体投影映射
- 基于语音和定位的视频自动标记
- 远程协作期间自动冻结实时视频以进行标注
- AVISA：视频理解标注工具
- BeaverDam：用于计算机视觉训练标签的视频标注工具
- BEDA：行为与生理传感器数据间关系的视觉分析
- CASAM：多媒体人机协作标注
- CLIPPER：物理学研究中的视听标注
- CoAT：基于网页的协作式标注工具
- CoVidA：基于笔的协作式视频标注
- 众包引导集成：如何协调众包工作者进行视频分割?
- CrowdSport：面向体育视频的基于众包的语义事件检测与视频标注
- DarkLabel
- 演示：语义人类活动标注工具——使用骨架化监控视频
- EagleView：用于可视化和查询人与设备之间空间交互的视频分析工具
- ELAN
- 利用众包为教程视频生成标注
- Glance：利用人群快速标注行为视频
- HistoryTracker：减少棒球比赛标注中的人机交互
- iSeg：视频中的半自动真实标注：用于基于多边形的对象标注和分割的交互式工具
- iVAT：用于人工、半自动和自动视频标注的交互式工具
- LabelMe
- Marquee：实时视频日志记录工具
- MediaDiver：查看和标注多视角视频
- MoViA：移动视频标注工具
- MRAS：网络流媒体视频标注
- 面向当代舞蹈创作的多模态视频标注
- MuLVAT：基于 XML词典和镜头聚类的视频标注工具

- NVivo
- Oudjat 致力于情绪面部表情（FEE）的手动标注
- 重新构想视频分析：交互式墨迹标注工具
- 通过社交与视觉协同标注重新思考在线新闻的参与方式
- Sirio、orione 和 pan：用于基于本体的视频搜索与标注的集成式网络系统
- 用于移动远程协助的稳定标注
- SVAT
- MESH 移动视频标注工具
- Timelinely
- Tool Eval：用于视频中目标检测的快速模型驱动标注与评估
- ToolScape：增强教程视频的学习体验
- VACA：用于定性视频分析的工具
- VAnnotator：将标注视为视频内容的多种视角
- VATIC
- VCode 和 VData：展示一种支持视频标注工作流程的新框架
- VIA：用于图像、音频和视频的 VIA 标注软件
- ViBAT
- VideoAnt
- VideoJot：多功能视频标注工具
- VidOR：用户生成视频中的物体和关系标注
- ViTBAT：视频跟踪与行为标注工具
- VoTT

C FEVA客户端架构

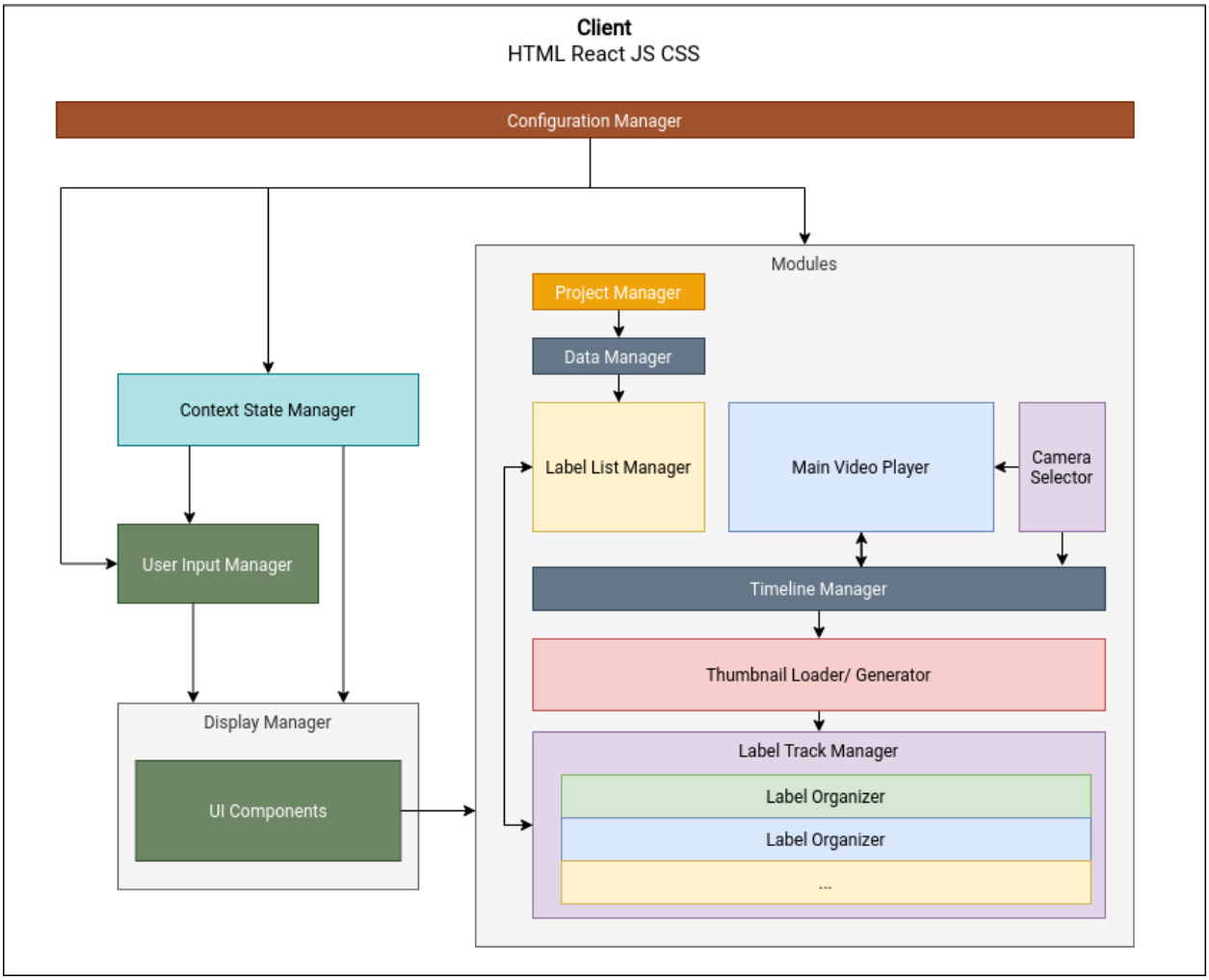


图8：FEVA客户端各模块的框图