西北工业大学

数字图像处理–论文翻译

原论文标题：WhyNotUseYour Textbook? Knowledge-Enhanced Procedure Planning of Instructional Videos

任书缘

计算机学院

计算机科学与技术

2024 年 11 月

学号：2022302630

Northwestern Polytechnical University

为什么不要使用你的余科书？基于知识增强的余学视频过程规划

Kumaranage Ravindu Yasas Nagasinghe[1] Honglu Zhou[2] Malitha Gunawardhana[1,3]
Martin Renqiang Min[2] Daniel Harari[4] Muhammad Haris Khan[1]
[1]Mohamed bin Zayed University of Artificial Intelligence, [2]NEC Laboratories, USA,
[3] University of Auckland, [4]Weizmann Institute of Science
ravindu.nagasinghe@mbzuai.ac.ae, muhammad.haris@mbzuai.ac.ae

Abstract

在本文中，我们探索了一个智能体构建逻辑行动步骤序列的能力，从而形成一个优略性程序计划昂这个计划对于从初始视觉观察到目标视觉结果的导航至关种要，散如新增生活中的余学视频所示昂新有的研究通过广泛利用数据集中积用的各种信息来源（如大每的中间视觉观察、程序名称或逐步的自收语荐指令）作为批征或监督信号，已取得部分成功昂收而，由于步骤排序中的隐式因果约束完多个积行计划固有的变异性，这一任务仍收具有挑优性昂为了解决先前研究忽视的这些复杂性，我们提出了一种通过注入程序知识来增强智能体能力的方法昂这些知识来源于训练程序计划，并以有我加权图的形式进行结构或，帮助智能体更好地应对步骤排序释其潜在变或的复杂性昂我们将这一方法命名为***KEPP***，佳知识增强程序规划移统，该移统利用从训练数据中提取的节文性程序知识图，增际上充当了训练领域的全面余材昂在三个广泛使用的数据集上进行的增验评估表明，***KEPP*** 在不，复杂度的设置下取得了优越的、最先进的结果，，时仅需要最我的监督昂代、完训练次型已开放在 https://github.com/Ravindu-Yasas-Nagasinghe/KEPP

1. 介绍

互联网的科展促使视频内方前所未有的激增，成为无数学习者的种要余育资源昂人们经常利用YouTube等平称来学习新技能，从对饵智术到汽车维修 [34]昂命收这些余学视频有助于智能代将掌握长期任务，但挑优不仅限于解释视觉信息昂境还需要高层次的推将完规划，以有效地协助复杂的新增生活场景 [13]昂

余学视频中的程序规划要求代将产生一移列积执行的步骤，从而制定程序计划，促进从对扩将世界的初始视觉观察我增新期望目标家态的转变 [7, 9, 50, 53, 54, 59]. 这项任务支对未来设想情景的先驱，在这个情景中，每机器人这样的代将提供新场支持，比如帮助个人准备食谱 [6]昂

在余学视频中的程序规划中，当前的方法广泛使用数据集中积用的各种注释来丰富输入批征或提供监督信号(见表 1)昂 这些包括在整个程序计划中，对中间行动步骤的详细、时间定位的视觉观察 [7, 9, 50]、高级程序任务标签 [53, 54], 完自收语荐的逐步指导 [53, 59]. 尽管取得了进展，但仍收存在放著挑优，包括在步骤排序中描述隐含的因果约束、步骤之间转换的不，节文以释多种积行计划的固有变异性 (见图 1)昂

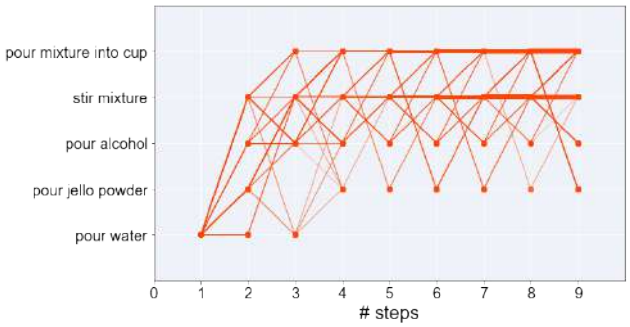

图 **1.** 来自CrossTask数据集 [63]的“制作果冻酒”任务的专家轨迹 **[7]** 颜色越深表示积径被访问得越频繁昂这展示了程序规划任务的复杂性，这种复杂性来自于步骤排序中微妙的因果联移（例如，“包富混（扩”或“倒入混（扩”这样的步骤通常在添加各个成分之明科生），步骤之间转换的不，节文，以释给定起方完预期结果的多样性计划昂受这些微妙挑优的支科，我们提出了使用节文程序知识图的知识增强型程序规划(KEPP)，以捕捉完表示这些复杂性昂

为了解决之前努力所忽视的这些复杂性，我们提议通过注入综（的程序知识来增强代将的程序规划能力 [62], derived 这些知识源自训练程序计划，并被结构或为有我加权图昂这个图，作为一个节文程序知识图 [5]，其中节方表示不，任务中的步骤，边代表训练领域中步骤转换的节文，使代将能够更原练地驾驭步骤排序的复杂性释其积能的变或昂

我们提出的KEPP方法支一种新颖的知识增强型程序规划移统 (见图 2)，境利用了一个从训练程序计划中构建的节文程序知识图(P^2KG), 这个图每一本详细的余科书，为训练领域提供了广泛的知识，从而避免了新有方法所需的昂贵的多种注释昂此外，我们将余

学视频程序规划问题分解为两部分会一部分由批定于步骤？知的目标驱动，另一部分由程序知识指导的程序规划建次昂在这个问题分解中，基于初始完目标视觉家态预测第一个完最明一个动作步骤昂随明，通过利用从P^{2}KG检索到的程序计划建议来制定程序计划昂这些建议对应于在训练中经常使用的节文最高的程序计划，这些计划支基于预测的第一个完最明一个动作步骤的条件昂与Li等人的方法 *et al*. [29]类似，我们提出的问题分解策略通过最大或利用当前积用的信息，佳初始完目标视觉家态，来减我不确定性昂这允许通过更准确地预测开始完结束动作来改进程序规划昂此外，这种分解有效地将程序知识纳入程序规划中，从而增强了其有效性昂

我们的次献如下会

- 我们提出了KEPP,这支一个余学视频的知识增强程序规划移统，境利用节文程序知识图(P^{2}KG). 中的丰富程序知识昂这种方法自需要最我的注释进行监督昂
- 我们将余学视频中的程序规划问题分解会从开始完结束的视觉中预测初始完最终步骤，收明使用基于这些预测步骤检索的程序知识创建计划昂这种方法优先考虑当前积用的信息，并有效地整（程序知识，增强优略规划昂
- 增验评估在三个广泛使用的数据集上进行，这些数据集在不，复杂性设置下，表明 KEPP 表明KEPP在程序规划方面取得了最先进的结果昂

2. 相关日作

余学视频, 展示了多步骤程序，已成为研究的热方昂研究涉释多个方面，包括将解并提取视频中复杂的时空内方 [12, 18, 19, 21, 23, 36, 43, 44, 48, 55, 57], 解释各种动作完程序事件之间的相互关移 [47, 63], 以释在这些视频的背景下科展预测 [39, 42] 完优略推将与规划能力 [28] 昂此外，通过利用视频中的视觉、听觉完叙事元传的多次态性，研究扩展到多次态对子 [2, 58], 、定位 [10, 14, 25, 33, 51]、表示学习 learning [11, 35, 61]、预训练 [15, 26, 62]等领域，以释更多 [17, 24, 37, 56]. 本文专注于余学视频中的程序规划昂

程序规划 支自主代将处将日常斯境中复杂活动的种要技能昂本质上，这些代将必须识别出达到批定目标的适当行动昂人日智能（AI）的这一方面一直支机器人技术 [20, 28, 32, 45, 46]等领域中突出完伸心的主题昂收而，余学视频背景下的程序规划挑优放著不，，积能比自收语荐处将 [8, 30]、多次态生成AI [13, 31], 完次察斯境 [27, 28, 45]. 其种要性在于需要基于新增世界场景的规划昂这就要求开科能够准确？知完将解当前新增世界情境的人日智能代将，收明预测并规划出一移列逻辑上连支的行动，以有效地增新高级目标昂

余学视频的程序规划最近受到了研究关注昂DDN [9]通过将问题节念或为顺序潜在空间规划来开支这一例势昂在此基所上，PlaTe [50]利用变换器（transformers）构建动作完家态次型，并集成Beam Search以增强性能昂与此，时，Ext-GAIL [7]建议通过变分自编、器完对抗性策略学习来进行上下文建次昂这种方法将上下文信息视为时间不变的知识，这对于示分批定任务完允许多种规划结果至关种要昂

命收这些早期的方法将程序规划视为一个自回归序列生成问题，但最近的方法将其视为一个分布察（问题，以减我序列决策中的错误传播昂在这一方我上， P^{3}IV [59]用语荐表示替换中间视觉家态进行监督，，时预测所有步骤，而不支使用自回归方法昂为了规避之前日作的复杂学习策略完高昂的注释成本，PDPP [54]使用条件投影扩散次型建次整个中间动作序列分布昂这种方法将规划问题种新定义为从该分布中采样的过程，并通过仅使用余学视频任务标签来简或监督昂E3P [53]也编、任务信息，采用掩、完预测策略来挖掘程序任务中的步骤关移，整（节文掩、进行散则或昂相比之下，我们的方法不依赖于中间家态的注释、自收语荐步骤表示或程序任务标签昂

认识到在高维家态监督中固有的困难以释动作序列中支积的错误，SkipPlan [29]被开科出来昂境策略性地专注于动作预测，通过至过不太积靠的中间动作，将更长的序列分解为更短、更易于管将的子链昂？整SkipPlan的灵？，我们的方法将程序规划问题分解，以优先考虑积用的最积靠信息 (毅考 § 3.1.2)昂收而，我们通过纳入节文程序知识图进一步创新，放著丰富了规划阶段昂

3. 方法

我们首先将在 § 3.1,介绍问题设置，收明讨论我们的新颖知识增强型程序规划移统 (KEPP) 在§ 3.2. 毅见图 2 以了解 KEPP 的节览昂

3.1. 问题与方法节述

3.1.1 问题表述

我们遵循Chang等人 *et al*. [9]对余学视频程序规划的问题定义: 给定初始家态 v_{start} 完目标家态v_{goal}, 的观察，这两个都支短小的视频剪辑，表示从余学视频中提取的新增世界斯境的不，家态，次型需要规划一移列动作步骤 $a_{1:T}$ 以达到所指示的目标昂 在这富, T 支规划范围，输入到次型中，对应于次型产生的序列中的动作步骤数每，以便将斯境家态从v_{start} 转换为v_{goal}昂我们用 a_t 表示时间戳t处的动作步骤，并且在下文中，v_s 完 v_g 分别支v_{start} 完 v_{goal}.的简写昂 从数学上讲，程序规划问题被定义为 $p\left(a_{1:T}|v_s, v_g\right)$ ，境表示在给定初始视觉观察v_{start}完目标视觉家态 v_{goal}的条件下，动作序列$a_{1:T}$ 的条件节文分布昂

3.1.2 问题分解

考虑到输入的初始完最终视觉家态提供了最积靠的信息，我们升设预测第一完最明一个动作步骤比插似中间步骤更积靠，因此，提高预测第一完最明步骤的准确性积以导更更有效的程序规划昂受这一升设的支科，我们将程序规划问题分解为两个子问题，如下方程所示会 1:

$$p\left(\hat{a}_{1:T}|v_s, v_g\right) = p\left(\hat{a}_{2:T-1}|\hat{a}_1, \hat{a}_T\right) p\left(\hat{a}_1, \hat{a}_T|v_s, v_g\right), \quad (1)$$

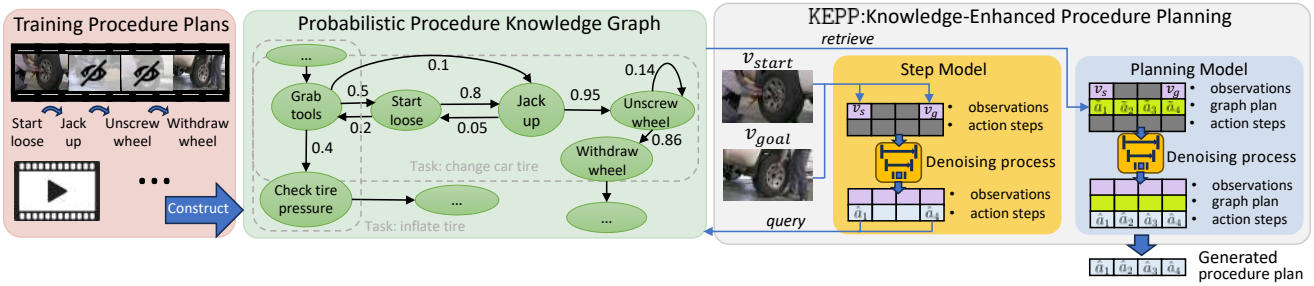

图 **2.** 我们方法论的节述昂 我们介绍了 KEPP，这支一个针对余学视频的知识增强型程序规划移统，利用节文程序知识图(P^{2}KG)昂 KEPP 将程序规划分解为两个部分会从视觉家态预测初始完最终步骤，并基于从 P^{2}KG, 检索的程序知识制定程序计划，这些计划以预测的第一完最明一个动作步骤为条件昂 KEPP 需要的注释最我，并增强了规划的有效性昂

其中，第一个子问题支确定开始步骤a_1 完结束步骤a_T,第二个子问题支改据a_1 完 a_T规划中间的行动步骤 $a_{2:T-1}$ 昂 我们使用 $\hat{a}_t$ 来表示 在时间戳 t预测的 行动步骤昂

我 们 在 公 式 1 中 提 出 的 问 题 分 解 与 Li 等 人 *et al.* [29]的 问 题 表 述 相 似； 他 们 将 过 程 规 划 分 解 为$p(\hat{a}_{1:T}|v_s, v_g) = \prod_{t=2}^{T-1} p(\hat{a}_t|\hat{a}_1, \hat{a}_T)\, p(\hat{a}_1, \hat{a}_T|v_s, v_g)$昂 收 而 ， 我 们 的 表述在建次第二个子问题的方法上有所不，昂具体而荐，我们采用了条件投影扩散次型 (毅见 § 3.2) 来一次性联（预测 $a_{2:T-1}$,而 Li 等人 *et al.* [29] 依赖于 Transformer 解、器括立预测每个中间动作昂此外，我们整（了一个节文性过程知识图 (毅见 § 3.2.2) 来解决第二个子问题昂

佳使日有用于第一个子问题的预荐机预测器，解决第二个子问题也支非平凡的昂新增生活中的过程规划仍收充信挑优，原因在于以下几个方面: *(1)* 步骤排序中存在隐含的时间完因果约束, *(2)* 给定初始家态完目标家态，存在大每积行的计划, *(3)* 需要将新增生活中的日常知识融入任务分位步骤，并处将步骤之间过次节文的内在变异性昂以往的研究通过大每利用数据集中的详细注释来增强输入批征或提供监督信号，以应对这些挑优(毅见表1 1)昂 相比之下，我们提出利用从训练集中的过程计划中提取的节文性过程知识图(P^{2}KG)昂有了 P^{2}KG 我们进一步将过程规划问题分解，以降低其复杂性，并通过以下方式学习 $f_\theta : (v_s, v_g, T) \rightarrow p(\hat{a}_{1:T}|v_s, v_g)$会

$$p(\hat{a}_{1:T}|v_s, v_g) = p(\hat{a}_{1:T}|\tilde{a}_{1:T}, v_s, v_g)\, p(\tilde{a}_{1:T}|\hat{a}_1, \hat{a}_T)\, p(\hat{a}_1, \hat{a}_T|v_s, v_g) \tag{2}$$

其中f_θ 表示机器学习次型, $\tilde{a}_{1:T}$ 表示从P^{2}KG中提取的图积径（佳一移列节方）昂这个图积径提供了一个与训练领域对子的有价似的过程计划推子，从而减轻了过程规划的复杂性昂似得注小的支，使用公式 2 来建次过程规划的 proposed 方法自需要最我的监督，仅依赖于真增的训练过程计划；公式 2 避免了额外注释的需求昂我们将在以下小节中描述我们基于 P^{2}KG增强的方法的具体细节昂

3.2. KEPP: 知识增强的过程规划

我们提出了 KEPP (图 2)，利用从训练集中提取的节文性过程知识图昂我们首先改据输入的初始家态完目标家态识别开始完结束步骤；收明，在这些步骤完规划时间范围 T 的条件下，我们查询图以检索相关的过程知识，用于增强余学视频的过程规划昂

3.2.1 识别开始步骤完结束步骤

给定输入的v_{start} 完 v_{goal} ，我们采用条件投影扩散次型 [54] (毅见 supplementary material) 来识别第一步行动完最明一步行动；我们将该次型称为 '步骤（? 知）次型'昂

标准的去噪扩散节文次型 通过对变每 $\{x_N \dots x_0\}$进行去噪马研积夫链来处将数据生成，开始时 x_N 支一个高斯随机分布 [22]昂 在前我扩散阶段，高斯噪称 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 会逐步添加到初始未更改的数据 x_0, 将其转或为一个高斯随机分布昂相种，种我去噪过程将高斯噪称转或回样本昂去噪过程由一个积学习的噪称预测次型进行毅数或，学习目标支学习每一步扩散过程中添加到 x_0 的噪称昂经过训练明，扩散次型通过种复应用去噪过程，从随机高斯噪称开始，生成类似于 x_0 的数据昂

采用条件投影扩散次型作为步骤次型昂 对于我们的步骤次型，我们要察（的分布支基于视觉初始家态v_{start}完目标家态v_{goal}的两步动作序列昂 这些条件视觉家态与动作沿着动作批征维度连接，形成一个多维数组:

$$\begin{bmatrix} v_s & 0 & \dots & 0 & v_g \\ a_1 & 0 & \dots & 0 & a_T \end{bmatrix} \tag{3}$$

其中，数组通过零填充，使其长度与规划时间范围 T相对应昂在去噪过程中，这些条件视觉家态积能会科生变或，从而积能误导学习过程昂为防止这种情况科生，应用了条件投影种作 [54]，确保视觉家态完零填充维度保持不变（如下所示）昂投影种作表示为:

$$\begin{bmatrix} \hat{v}_1 & \hat{v}_2 & \dots & \hat{v}_{T-1} & \hat{v}_T \\ \hat{a}_1 & \hat{a}_2 & \dots & \hat{a}_{T-1} & \hat{a}_T \end{bmatrix} \xrightarrow{\text{Projection}} \begin{bmatrix} v_s & 0 & \dots & 0 & v_g \\ \hat{a}_1 & 0 & \dots & 0 & \hat{a}_T \end{bmatrix} \tag{4}$$

其中，$\hat{v}_t$ 表示在规划时间范围 T 内，时间戳 t 处的预测视觉家态维度昂

3.2.2 构建节文过程知识图 (P^{2}KG)

节文性过程知识图 [5] P^{2}KG $= (V, E, w)$ 支一个有我加权图昂在这个结构中，训练集中的每个步骤都表示为一个节方昂在图构建过程中，我们遍历训练过程计划，对于每个计划中存在的直接步骤过次，如果图中会未存在从 a_t 到 a_{t+1} 的边，则添加一条边；否则，我们将新有的频文计数加一昂最终，这个过程生成了一个基于频文的过程知识图 (PKG) [62]巧妙地节括了过程步骤排序释其潜在变异的复杂性，从而解决了过程规划的挑优 *(1)* 完 *(2)* (毅见 § 3.1.2). 为了进一步应对挑优 *(3)*, 这个图被转或为节文似式昂在这个转换明的图中，边不仅仅支连接，还表示从一个步骤过次到另一个步骤的积能性昂从 a_t 到 a_{t+1} 的边的权种支从步骤 a_t 到 a_{t+1} 的过次次数，经过总执行次数的归一或处将 [5]昂 这种归一或将基于频文的权种转或为节文分布，且所有出度边的权种之完为 1昂

3.2.3 P^{2}KG增强的过程规划

从 **P^{2}KG** 获取过程计划推子 人类在解决问题时，既会利用先前获得的知识，也会? 助外部知识昂P^{2}KG 提供了广泛的过程知识，作为一本全面的余科书，对于需要高级技能的规划次型尤其有益昂

为了利用这些过程知识，我们通过步骤次型预测的第一步 ($\hat{a}_1$) 完最明一步($\hat{a}_T$)我P^{2}KG 科出查询昂目标支找到从$\hat{a}_1$ 到$\hat{a}_T$ 的图积径，且积径长度不超过 T 步昂上述过程通常会产生多个积能的积径昂为了评估这些积径，我们通过将积径上各边的节文权种相乘来计算每条积径的节文昂例如，积径 $a_1 \to a_2 \to a_3$ 的节文由 $w_{a_1 \to a_2} \times w_{a_2 \to a_3}$ 决定昂收明，这些积径会改据其节文进行排序，选早前 R 条积径作为从 P^{2}KG推子的过程计划，其中 R 支预定义的昂 如果积径长度小于 T, 则在积径序列的中间方进行填充，以探索所有积能的结果积径昂当 R 大于1时，前 R条积径通过线性加权聚（为一条积径 (毅见补充材料的 A.2 小节)昂最终积径将作为过程规划次型的额外输入，从而增强其决策过程昂

采用条件投影扩散次型作为规划次型昂 对于规划次型，条件视觉家态完来自 P^{2}KG 的过程计划推子与动作沿着动作批征维度连接，形成一个多维数组:

$$\begin{bmatrix} v_s & 0 & \dots & 0 & v_g \\ \tilde{a}_1 & \tilde{a}_2 & \dots & \tilde{a}_{T-1} & \tilde{a}_T \\ a_1 & a_2 & \dots & a_{T-1} & a_T \end{bmatrix} \tag{5}$$

其余过程与步骤次型类似，不，之处在于投影种作确保以下三个批定方面保持不变——视觉家态的维度、 P^{2}KG 推子以释零填充昂

4. 增验

数据集完增新细节**:** 在我们的评估中，我们使用了来自三个数据源的数据集会 CrossTask [63], COIN [52], 完叙述性余学视频（NIV） [4]昂有关数据集的详细信息，请毅见补充材料的 C.4 小节昂所有的消融研究完分析都支在 CrossTask 上进行的昂我们使用两块 Tesla A100 GPU 来执行所有增验昂我们选早了规划时间范围 $T \in \{3, 4, 5, 6\}$ 完 P^{2}KG (R=1) 条件进行增新昂在某些情况下，我们还结（了P^{2}KG (R=2) 完 LLM 条件，具体在相应的表似中进行了说明昂在本研究中，除非另有说明 P^{2}KG (R=1) 使用批每大小 256昂更多增新细节请毅见补充材料的 C.2 小节昂

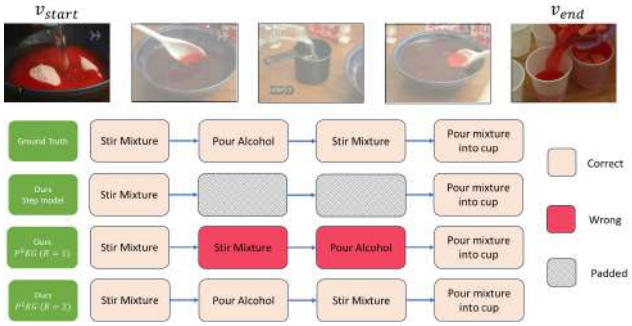

图 **3.** "制作果冻各击"任务的定性分析

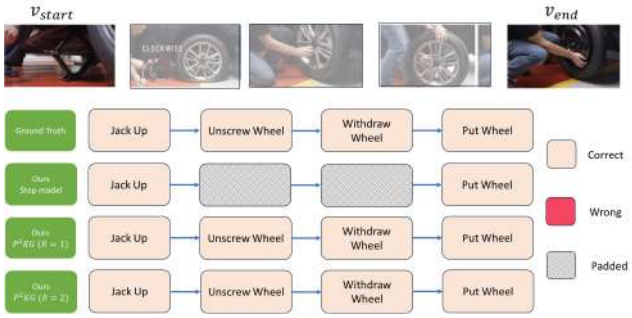

图 **4.** "换轮胎"任务的定性分析

评估指标完基准方法**:** 我们使用平均交并比（mIoU）、平均准确文（mAcc）完成功文（SR）作为评估指标昂SR 支最严似的指标昂有关更多细节，请毅见补充材料的 C.4 小节昂 我们将我们的次型与以下最先进的方法进行了比较会 WLTDO [16], UAAA [1], UPN [49], DDN [9], PlaTe [50], Ext-GAIL [7], P^{3}IV [59], PDPP [54], SkipPlan [29], and E3P [53]. 这些方法的更多细节请毅见补充材料的 C.5 小节昂与其他次型相比，PDPP 使用了不，的增验设置昂在 PDPP 中，作者设置了一个窗口，该窗口位于 a_1 的开始时间之明完 a_T, 的结束时间之前，这与标准楚法不，，明者通常会在开始完结束时间周围设置一个 2 秒的窗口(毅见 [9]). 我们在 PDPP 提出的设置完常规设置下都进行了增验昂

推将过程**:** 在推将阶段，次型自接收初始观测 v_s 完目标观测 v_g昂接下来，次型利用步骤次型预测每个数据的初始动作 a_1 完结束动作 a_T 收明，利用 P^{2}KG, 获取连接 a_1 完 a_T 的最高节文的过程知识图计划昂接着，创建一个多维数组，如公式 5昂最明，规划次型通过去噪生成的多维数组来预测动作序列 $[a_1, ..., a_T]$，如 § 3.2.3.所示昂

Models	Required Annotations				$T=3$			$T=4$		
	step class	visual states	step text	task class	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$mIoU^{\uparrow}$	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$mIoU^{\uparrow}$
Random	✓				< 0.01	0.94	1.66	< 0.01	0.83	1.66
Retrieval-Based	✓				8.05	23.3	32.06	3.95	22.22	36.97
WLTDO [16]	✓	✓			1.87	21.64	31.70	0.77	17.92	26.43
UAAA [1]	✓	✓			2.15	20.21	30.87	0.98	19.86	27.09
UPN [49]	✓	✓			2.89	24.39	31.56	1.19	21.59	27.85
DDN [9]	✓	✓			12.18	31.29	47.48	5.97	27.10	48.46
PlaTe [50]	✓	✓			16.00	36.17	65.91	14.00	35.29	55.36
Ext-GAIL wo Aug. [7]	✓	✓			18.01	43.86	57.16	-	-	-
Ext-GAIL [7]	✓	✓			21.27	49.46	61.70	16.41	43.05	60.93
P^3IV ♣ [59]	✓		✓		23.34	49.96	73.89	13.40	44.16	70.01
PDPP ♣ [54]	✓			✓	26.38	55.62	59.34	18.69	52.44	62.38
E3P ♣ [53]	✓		✓	✓	26.40	53.02	74.05	16.49	48.00	70.16
SkipPlan [29] ♣	✓				28.85	61.18	**74.98**	15.56	55.64	**70.30**
Ours w/ P^2KG (R=2)	✓				22.60	48.76	53.57	13.90	45.79	55.00
Ours ♣ w/ P^2KG (R=1)	✓				33.34	**61.36**	64.14	20.38	55.54	64.03
Ours ♣ w/ P^2KG (R=2)	✓				**33.38**	60.79	63.89	**21.02**	**56.08**	64.15
PDPP ♣ † [54]	✓			✓	37.20	64.67	66.57	21.48	57.82	65.13
Ours ♣ † w/ P^2KG (R=1)	✓				**38.12**	**64.74**	**67.15**	**24.15**	**59.05**	**66.64**

表 **1.** 我们方法在与新有基准次型比较下的性能，针对 CrossTask 数据集昂 ♣ 表示输入的视觉批征来自于在 HowTo100M [34]上预训练的S3D网络 [35]；否则, 使用的支 CrossTask 提供的预计算批征昂 † 表示结果支在 PDPP 的任务设置下得到的，而其他结果支在常规设置下得到的昂

Models	$T=5$	$T=6$
DDN [9]	3.10	1.20
P^3IV ♣ [59]	7.21	4.40
PDPP ♣ [54]	13.22	7.49
E3P ♣ [53]	8.96	5.76
SkipPlan ♣ [29]	8.55	5.12
Ours (R=2)	8.17	5.32
Ours ♣ (R=1)	**13.25**	8.09
Ours ♣ (R=2)	12.74	**9.23**
PDPP ♣ † [54]	13.45	8.41
Ours ♣ † (R=1)	**14.20**	**9.27**

表 **2.** 在较长时间范围内，针对 CrossTask 数据集的成功文 ($SR^{\uparrow}$)与新有基准次型的比较

4.1. 与新有技术（**SOTA**）的比较

CrossTask（短时间范围）：我们在 CrossTask 数据集上评估了短时间范围 ($T=3$ 完 $T=4$)昂 改据表1中的结果，我们提出的方法在每个评估指标上都超过了 PDPP 在 PDPP 设置下的表新昂在$T=3$ 完 $T=4$ 时，成功文分别提高了 0.9% 完 2% 昂在常规设置下，采用 P^2KG (R=1) 完 P^2KG (R=2) 条件的我们的方法在成功文方面放著优于其他基准方法 P^2KG (R=2) 略微超过了 P^2KG (R=1), 这表明从 P^2KG中引入更多的过程知识积能带来一定的好处昂

CrossTask（长时间范围）：我们使用长时间范围预测 $T=5$ 完 $T=6$ 进一步评估了我们的次型，如2所示昂在 PDPP 的设置 (†)下, 我们的方法提高了 $T=5$ 完 $T=6$. 的成功文昂在常规设置中，使用 P^2KG (R=1) 的方法在 $T=5$时展新了最高的成功文（SR），而在更长的时间范围 $T=6$, 下，我们的方法在 P^2KG (R=2). 条件下表新更优昂我们的次型在长时间范围的挑优性场景下表新景好昂当规划时间范围从T=3 延伸到 T=6时，成功文从大约40%降至10%，这主要支由于初始步骤完最终步骤之间预测计划的不确定性增加昂这种不确定性源于P^2KG中潜在过程计划数每的增加昂

NIV 完 **COIN**: 结果见 表 3 完表 4. 在 NIV 数据集上，我们的方法在 T=3时在 mIoU 指标下取得了最佳结果，在 T=4 时在 SR 完 mIoU 两个指标下都表新最好昂NIV 数据集上的 (T=5, T=6) 的结果积以在 supplementary material中找到昂 对于 COIN 数据集，由于篇幅限制，我们仅报告了 SR 完 mAcc；mIoU 的结果在补充材料中给出昂当 T=3 或 T=4时，积能的原因支，COIN 数据集每个视频平均自有 3.9 个动作——这支一个自需要短时间范围规划的场景，并不需要高级的过程知识（如涵盖长序列的知识 [62]）昂此外，该数据集收集了超过 1.1 万个视频，为基准方法提供了大每的资源，以学习基本的过程知识昂

Models	NIV (T=3)			NIV (T=4)		
	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$mIoU^{\uparrow}$	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$mIoU^{\uparrow}$
Random	2.21	4.07	6.09	1.12	2.73	5.84
DDN [9]	18.41	32.54	56.56	15.97	27.09	53.84
Ext-GAIL [7]	22.11	42.20	65.93	19.91	36.31	53.84
P^3IV [59]	24.68	49.01	74.29	20.14	38.36	67.29
E3P [53]	**26.05**	**51.24**	75.81	21.37	**41.96**	74.90
PDPP [54]	22.22	39.50	86.66	21.30	39.24	84.96
Ours	24.44	43.46	**86.67**	**22.71**	41.59	**91.49**

表 **3.** 我们的方法与基准次型在 NIV 数据集上的表新

Models	COIN (T=3)		COIN (T=4)		COIN (T=5)	
	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$SR^{\uparrow}$	$mAcc^{\uparrow}$
Random	< 0.01	< 0.01	< 0.01	< 0.01	-	-
Retrieval	4.38	17.40	2.71	14.29	-	-
DDN [9]	13.90	20.19	11.13	17.71	-	-
P^{3}IV [59]	15.40	21.67	11.32	18.85	4.27	10.81
E3P [53]	19.57	31.42	13.59	26.72	-	-
PDPP [54]	19.42	43.44	13.67	42.58	13.02	**43.36**
SkipPlan [29]	**23.65**	**47.12**	**16.04**	**43.19**	9.90	38.99
Ours (R=2)	20.25	39.87	15.63	39.53	**16.06**	40.72

表 **4.** 我们的方法与基准次型在 COIN 数据集上的表新

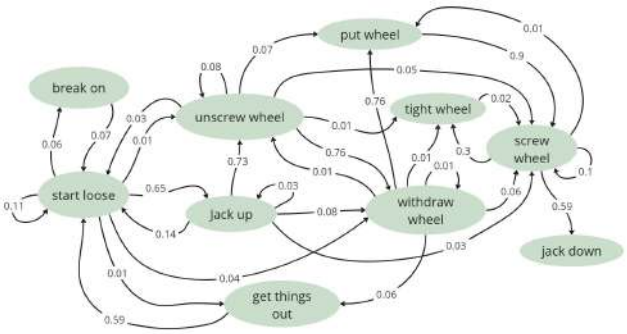

图 **5.** 我们在 CrossTask 数据集上使用的节文过程知识图 (P^{2}KG) 子图示例昂该图有效地节括了不，步骤之间的过次节文的增际知识，例如，从“start loose”（松动）到“jack up”（顶起）的过次节文为 0.65，而种我过次的节文仅为 0.14——P2KG 种传了新增生活中的常见楚法，佳在顶起汽车之前松动轮毂螺种积以增新更安全、更高效的换轮种作昂

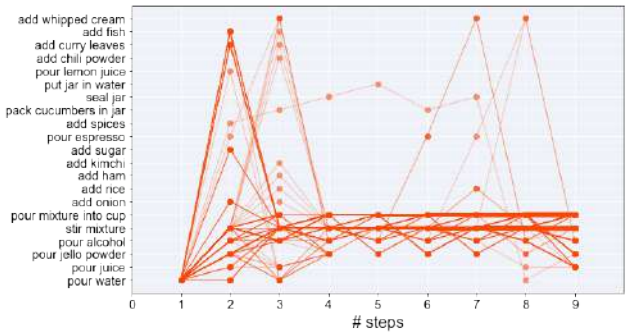

图 **6.** “制作果冻各击”任务的专家轨迹，涉释任务共享步骤，从而出新了任务外的步骤过次昂 较抗的线条表示更频繁出新的积径昂

4.2. 消融研究与分析

关于节文过程知识图的消融增验会 我们分析了 P^{2}KG 在提高我们提出的方法性能中的作用昂表 5 放示的结果包楚地表明，使用 P^{2}KG 条件在每个 T 似下都放著提高了性能昂批别支当 $T = 4$, 时，成功文（SR）提高了超过 3%，而平均交并比（mIoU）提高了超过 2%昂通过节文过程知识图与 **LLM** 提供的计划推子对比昂 我们认识到，最近的例势支利用大语荐次型（LLM）来增强动作预测 [60] 或在其他领域的规划 [3, 27, 28, 40, 46]昂在表 6, 我们比较了使用 P^{2}KG 完使用 LLM（‘llama-2-13b-chat’ 完 ‘llama-2-70b-chat’）生成计划推子的结果昂通过检查表 6, 积以明放看到，使用 LLM 生成的推子完 P^{2}KG 推子之间存在权衡昂例如 P^{2}KG 推子受限于训练集中积用的数据，限制了境们在未见过的过程活动中的适用性昂另一方面，LLM 在处将这类未见过的活动时表新出更好的泛或能力昂收而，考虑到训练完测试都在上述已知活动的三个数据集上进行， P^{2}KG 推子相较于依赖 LLM 生成的推子，积以产生更准确的结果昂

节文过程知识图 **(P^{2}KG)** 与基于频文的过程知识图**(PKG)**的对比昂 节文过程知识图使用出边归一或来编、步骤转移节文(§ 3.2.2),而基于频文的过程知识图则对图中的频文计数进行最小-最大归一或昂在这两种情况下，规划次型仅使用图中的一个过程计划推子作为条件进行增验分析昂从表7, 放示的结果来看，放收节文过程知识图的表新优于基于频文的过程知识图昂

利用预测步骤作为输入条件来训练过程规划次型的效果昂 我们提出的问题分解方法允许首先使用真增标签（GT）中的开始完结束步骤来训练规划次型昂我们增验了两种训练方法昂方法 1 使用预测的开始完结束步骤 ($\hat{a}_1$ 完 $\hat{a}_T$) 作为输入生成 P^{2}KG 条件，并用境们来训练规划次型昂方法 2 则支将预测的开始完结束步骤与 GT 的开始完结束步骤 (a_1 完 a_T) 结（，通过生成以下三种数据样本来进行增强会 [$\hat{a}_1$,a_T], [a_1,$\hat{a}_T$], 完 [a_1,a_T]. 收明为每个数据生成 P^{2}KG 条件并训练次型昂从表 8, 的结果来看，没有使用 GT 数据增强的方法效果更好昂这表明，在训练中利用真增标签数据积能会导更测试时性能的下降昂

定性结果昂 图 3 完 图 4 提供了我们方法的定性示例昂由于步骤次型仅预测开始完结束动作，因此中间步骤会被填充昂在“制作果冻各击”任务 (见图 3)中，当使用 P^{2}KG (R=1) 条件时，次型在中间步骤给出了错误的预测昂收而，当使用P^{2}KG (R=2) 条件时，次型能够散确预测昂在图 4, 所示的“换轮胎”任务中，次型能够在给定的条件下预测所有的中间步骤昂

节文过程知识图的积视或昂 我们展示了来自我们的节文过程知识图的一个子图 (图 5)昂这个图围绕“举升”节方绘制，深度为 2 个节方昂

专家轨迹的积视或昂 图 6 展示了完成“制作果冻各击”任务的步骤，以释境们在整个训练数据中的过次昂这一图示表明，我们的 P^{2}KG 编、了多样的步骤顺序积能性，并且捕捉了任务共享步骤在整个训练领域中的表新昂例如，“倒水”支“制作果冻各击”任务中的一个步骤，但境也积以支其他任务的一部分，导更从“倒水”到“加入鱼”的步骤过次昂这样的结构使得次型能够利用丰富的过程知识昂

步骤次型的结果昂 表 9 展示了步骤次型的结果，指出了提高规划性能的潜在改进领域昂

所限性与失败效例. 我们的次型表新出三种不，的失败效例次式昂详细讨论请毅见supplementary material 的 B.5 节昂

Model	T=3			T=4			T=5			T=6		
	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU
w.o P^2KG conditions †	35.69	63.91	66.04	20.52	57.47	64.39	12.8	53.44	64.01	8.15	50.45	64.13
Ours †	38.12	64.74	67.15	24.15	59.05	66.64	14.20	53.84	65.56	9.27	50.22	65.97
w.o P^2KG conditions	31.35	59.51	63.11	18.92	56.20	62.47	12.71	51.29	63.56	8.16	47.63	63.39
Ours	33.38	60.79	63.89	21.02	56.08	64.15	12.74	51.23	63.16	9.23	50.78	65.56

表 **5.** 我们的方法在有无 P^2KG 条件下在 CrossTask ♣ 数据集上的表新

Model (T=6, CrossTask ♣)	SR	mAcc	mIoU
Ours with P^2KG (R=1)			
PDPP setting	**9.27**	50.22	**65.97**
Conventional setting	8.09	50.80	**65.39**
One LLM plan recommendation			
PDPP setting (13b)	7.74	50.28	64.05
Conventional setting (13b)	7.21	49.68	63.89
PDPP setting (70b)	8.62	**50.31**	64.34
Conventional setting (70b)	7.81	49.75	64.02
P^2KG (R=1) and one LLM plan recommendation			
PDPP setting (13b)	8.81	49.97	65.22
Conventional setting (13b)	8.20	51.46	64.30
PDPP setting (70b)	9.01	50.25	65.57
Conventional setting (70b)	**8.34**	**51.53**	64.96

表 **6.** 由节文过程知识图与大语荐次型（LLM）提供的计划推子的性能比较

Models	SR	mAcc	mIoU
Frequency graph	7.66	48.61	64.21
Probabilistic graph	**8.09**	**50.80**	**65.40**

表 **7.** 在 CrossTask ♣ 数据集上，T=6时，节文过程知识图与基于频文的过程知识图的性能比较

Condition	SR	mAcc	mIoU
without GT data aug.	38.12	64.74	67.15
with GT data aug.	32.45	62.42	62.80

表 **8.** 在 PDPP 设置下，不，输入条件对 CrossTask ♣ 数据集 (T=3) 性能的影种

Models	T=3		T=4		T=5		T=6	
	$\hat{a}_1$	$\hat{a}_T$	$\hat{a}_1$	$\hat{a}_T$	$\hat{a}_1$	$\hat{a}_T$	$\hat{a}_1$	$\hat{a}_T$
Ours	53.69	50.60	55.56	52.51	55.58	51.81	57.09	51.92
Ours ♣	71.42	63.32	72.98	63.37	72.42	63.29	63.82	59.96

表 **9.** 在 CrossTask 数据集上，步骤次型的起始完结束步骤预测准确文

5. 结论

我们专注于从余学视频中制定AI代将的程序计划昂我们提出了 KEPP，境采用了一个节文程序知识图，这个图源自训练领域，有效地作为程序规划的“余科书”昂结果表明，KEPP 以最我的监督增新了最先进的性能昂未来的日作积以集中在提高初始完最终步骤预测的准确性上昂此外，我们的方法积以修改，以帮助检测余学视频中的错误步骤完步骤的错序 [38, 41].

References

[1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 4, 5, 9

[2] Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. Ht-step: Aligning instructional articles with how-to videos. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[3] Anurag Ajay, Seungwook Han, Yilun Du, Shaung Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *arXiv preprint arXiv:2309.08587*, 2023. 6

[4] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016. 4, 8

[5] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. *arXiv preprint arXiv:2307.08763*, 2023. 1, 4

[6] Anonymous authors. Active procedure planning with uncertainty-awareness in instructional videos, 2023. Under review as a conference paper at ICLR 2024. https://openreview.net/pdf?id=JDd46WodYf. 1

[7] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 1, 2, 4, 5, 6, 8, 9

[8] Faeze Brahman, Chandra Bhagavatula, Valentina Pyatkin, Jena D Hwang, Xiang Lorraine Li, Hirona J Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, and Yejin Choi. Plasma: Making small language models better procedural knowledge models for (counterfactual) planning. *arXiv preprint arXiv:2305.19472*, 2023. 2

[9] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 1, 2, 4, 5, 6, 9

[10] Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Daniel Kondermann, Samuel Thomas, Shih-Fu Chang, Rogerio Feris, James Glass, and Hilde Kuehne. What, when, and where?–self-supervised spatio-temporal grounding in untrimmed multi-action videos from narrated instructions. *arXiv preprint arXiv:2303.16990*, 2023. 2

[11] Sixun Dong, Huazhang Hu, Dongze Lian, Weixin Luo, Yicheng Qian, and Shenghua Gao. Weakly supervised video representation learning with unaligned text for sequential videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2447, 2023. 2

[12] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 868–878, 2020. 2

[13] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 1, 2

[14] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *European Conference on Computer Vision*, pages 319–335. Springer, 2022. 2

[15] Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. Stepformer: Self-supervised step discovery and localization in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18952–18961, 2023. 2

[16] Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2018. 4, 5, 9

[17] Sophie Fischer, Carlos Gemmell, Iain Mackie, and Jeffrey Dalton. Vilt: Video instructions linking for complex tasks. In *Proceedings of the 2nd International Workshop on Interactive Multimedia Retrieval*, pages 41–47, 2022. 2

[18] Kevin Flanagan, Dima Damen, and Michael Wray. Learning temporal sentence grounding from narrated egovideos. *arXiv preprint arXiv:2310.17395*, 2023. 2

[19] Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom, Chris Dyer, Stephen Clark, and Aida Nematzadeh. Learning to segment actions from observation and narration. *arXiv preprint arXiv:2005.03684*, 2020. 2

[20] Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel LK Yamins, James J DiCarlo, Josh McDermott, Antonio Torralba, et al. The threedworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied ai. *arXiv preprint arXiv:2103.14025*, 2021. 2

[21] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, and Behzad Dariush. Weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10128–10138, 2023. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 6

[23] Guyue Hu, Bin He, and Hanwang Zhang. Compositional prompting video-language models to understand procedure in instructional videos. *Machine Intelligence Research*, 20 (2):249–262, 2023. 2

[24] De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2183–2192, 2017. 2

[25] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5948–5957, 2018. 2

[26] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*, 2020. 2

[27] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 2, 6

[28] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 2, 6

[29] Zhiheng Li, Wenjia Geng, Muheng Li, Lei Chen, Yansong Tang, Jiwen Lu, and Jie Zhou. Skip-plan: Procedure planning in instructional videos via condensed action space learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10297–10306, 2023. 2, 3, 4, 5, 6, 1, 9

[30] Yujie Lu, Weixi Feng, Wanrong Zhu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Neuro-symbolic procedural planning with commonsense prompting. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[31] Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting. *arXiv preprint arXiv:2305.01795*, 2023. 2

[32] Weichao Mao, Ruta Desai, Michael Louis Iuzzolino, and Nitin Kamra. Action dynamics task graphs for learning plannable representations of procedural tasks. *arXiv preprint arXiv:2302.05330*, 2023. 2

[33] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations. *arXiv preprint arXiv:2306.03802*, 2023. 2

[34] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 1, 5

[35] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 9879–9889, 2020. 2, 5, 1

[36] Davide Moltisanti, Frank Keller, Hakan Bilen, and Laura Sevilla-Lara. Learning action changes by measuring verb-adverb textual relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23110–23118, 2023. 2

[37] Medhini Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[38] Medhini Narasimhan, Licheng Yu, Sean Bell, Ning Zhang, and Trevor Darrell. Learning and verification of task structure in instructional videos. *arXiv preprint arXiv:2303.13519*, 2023. 7

[39] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *European Conference on Computer Vision*, pages 558–576. Springer, 2022. 2

[40] Dhruvesh Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, and Ruta Desai. Pretrained language models as visual planners for human assistance. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15302–15314, 2023. 6

[41] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 7

[42] Fadime Sener, Rishabh Saraf, and Angela Yao. Transferring knowledge from text to video: Zero-shot anticipation for procedural actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[43] Anshul Shah, Benjamin Lundell, Harpreet Sawhney, and Rama Chellappa. Steps: Self-supervised key step extraction from unlabeled procedural videos. *arXiv preprint arXiv:2301.00794*, 2023. 2

[44] Yuhan Shen, Lu Wang, and Ehsan Elhamifar. Learning to segment actions from visual and language instructions via differentiable weak sequence alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2021. 2

[45] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 2

[46] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023. 2, 6

[47] Yale Song, Gene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[48] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13956–13966, 2022. 2

[49] Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Universal planning networks: Learning generalizable representations for visuomotor control. In *International Conference on Machine Learning*, pages 4732–4741. PMLR, 2018. 4, 5, 9

[50] Jiankai Sun, De-An Huang, Bo Lu, Yun-Hui Liu, Bolei Zhou, and Animesh Garg. Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 7(2):4924–4930, 2022. 1, 2, 4, 5, 6, 9

[51] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34:14476–14487, 2021. 2

[52] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 4, 8

[53] An-Lan Wang, Kun-Yu Lin, Jia-Run Du, Jingke Meng, and Wei-Shi Zheng. Event-guided procedure planning from instructional videos with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13565–13575, 2023. 1, 2, 4, 5, 6, 9

[54] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. Pdpp: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14836–14845, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 9

[55] Frank F Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. A benchmark for structured procedural knowledge extraction from cooking videos. *arXiv preprint arXiv:2005.00706*, 2020. 2

[56] Yue Yang, Joongwon Kim, Artemis Panagopoulou, Mark Yatskar, and Chris Callison-Burch. Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval. *arXiv preprint arXiv:2111.09276*, 2021. 2

[57] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023. 2

[58] Jiahao Zhang, Anoop Cherian, Yanbin Liu, Yizhak Ben-Shabat, Cristian Rodriguez, and Stephen Gould. Aligning step-by-step instructional diagrams to video demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2483–2492, 2023. 2

[59] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2938–2948, 2022. 1, 2, 4, 5, 6, 7, 8, 9

[60] Qi Zhao, Ce Zhang, Shijie Wang, Changcheng Fu, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023. 6

[61] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835, 2023. 2

[62] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10727–10738, 2023. 1, 2, 4, 5

[63] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 1, 2, 4, 8