arXiv:2301.02311v2 [cs.CV] 8 Jun 2023

HierVL：学习层次视频-语言嵌入

Kumar Ashutosh[1], Rohit Girdhar[2],Lorenzo Torresani[2],Kristen Grauman[1,21]德克萨斯大学奥斯汀分校， [2]FAIR，Meta AI

摘要

视频语言嵌入是将语义注入视觉表征的一种有前景的途径，但现有方法只能捕捉持续数秒的视频片段与其配套文本之间的短期关联。我们提出了 *HierVL*，这是一种新颖的分层视频-语言嵌入，能够同时考虑长期和短期关联。作为训练数据，我们采用附有描述人类动作的带时间戳文本描述的视频，以及概括长视频中整体活动的高层文本摘要（例如 *Ego4D* 中提供的数据）。我们提出了一种层次对比训练目标，鼓励在片段级和视频级实现文本-视觉对齐。片段级约束利用逐步描述捕捉那一瞬间正在发生的事情，而视频级约束则利用摘要文本捕捉事情发生的原因，即活动的更广泛背景以及行动者的意图。我们的层次方案生成的片段表征优于其单层对应方法，同时生成的长期视频表征在需要长期视频建模的任务上取得了最先进水平的结果。*HierVL* 在零样本和微调设置下，均成功迁移至多个具有挑战性的下游任务（*EPIC-KITCHENS-100*、*Charades-Ego*、*HowTo100M*）。

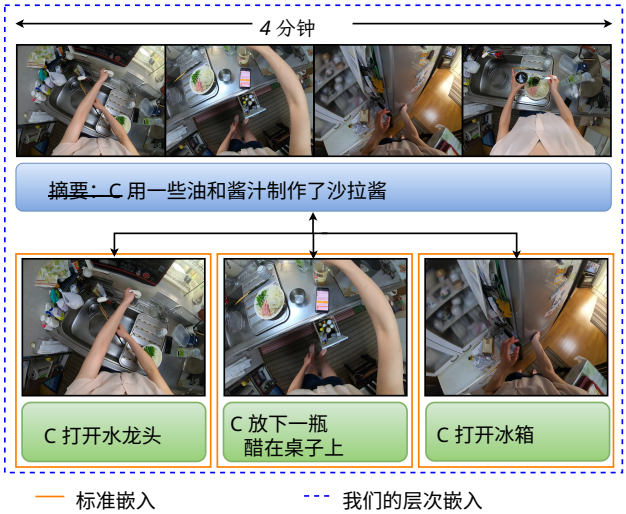


图 1. 传统的视频-语言嵌入经过训练，将短时片段与其对应的描述相匹配，例如打开水龙头（橙色框），从而捕捉正在发生什么。我们的分层视频-语言嵌入（蓝色虚线框）能够学习短期和长期的视觉-文本关系，进而捕捉事情为何发生（例如，制作沙拉酱）。长期意图通过文本摘要（蓝色）传达，这些摘要对整个视频进行了抽象概括，并与更为具体的逐步叙述（绿色）相辅相成。

1. 引言

视频中的人类活动理解是一项基础的视觉问题，在增强现实、机器人学和信息检索等领域有着广泛的应用。该领域取得了令人振奋的进展，从新的识别模型 [24,53, 86] 和自监督表征 [55, 58, 61, 90]到大型数据集 [16, 34, 63, 74, 106]。尽管如此，视频中的活动理解仍明显落后于图像中的物体理解；如今的人工智能模型在物体理解方面已能与人类一较高下。

造成这种差异的一个关键原因是：物体直接呈现在像素中——无需任何潜台词——而活动天然具有广泛的时间背景，根植于人类行动者的（潜在）意图。活动不仅会跨越多个视频帧，其解读还取决于更广泛的背景，即这个人试图完成什么。因此，视频中的信息天然具有层次结构，从短期的“这个人此刻正在做什么”（例如伸手去拿炉子上的东西），一直延伸到长期的“这个人想要做什么”（例如做晚餐）。

为了捕捉这种层次结构，我们探索了视频-语言表征学习。视频通常伴有带时间戳的文本，无论这些文本来自教程视频中的口述解说 [63, 75, 106], 隐藏式字幕文本和脚本 [9,76], , 还是有意添加的文本标注 [16,34,91]。现有的视频-语言模型通过将短视频片段与对应文本进行匹配，学习这两种模态之间的对应关系，通常会学习一个嵌入

网站：https://vision.cs.utexas.edu/projects/hiervl/

向量 [3, 55, 61, 90]，以生成语言增强型视频片段编码器。然而，这种标准方法可能只捕捉到短期动作。诸如“现在我把牛奶倒进锅里”或“他拿起了一根水管”这样的细致评论，无法捕捉活动的整体目标，例如煮咖啡或洗车。因此，在推理时，对于未见过的视频，这些模型生成的编码可能会目光短浅，无法捕捉观测到的事件之间的序列依赖关系。

为解决这一问题，我们提出 HierVL：一种新颖的层次化视频语言模型，能够捕捉视频中的短期动作和长期意图。不同于标准的视频-语言嵌入，我们的方法旨在同时捕捉即时观察到的动作，以及这些动作对长期目标的贡献。为此，给定带有时间戳片段级文本描述和全局（视频级）文本摘要的训练视频，HierVL 通过两层对比学习，学习用于层次化时间理解的视频文本嵌入。顶层（父层）促使聚合后的视频片段接近整体文本摘要（例如，他做了一顿意大利面晚餐），而底层（子层）则训练单个片段，使其与相应的描述相似（例如，他打开炉灶）。见图1。

据我们所知，我们是首个构建层次化视频-语言嵌入的工作。将抽象的文本摘要与直观的文本描述相结合，是我们的新颖构想。此外，我们的模型设计解决了若干关键技术挑战——具体而言，我们通过聚合短期特征，避免了长期特征学习通常所需的高昂成本 [4, 43, 86]，并展示了如何联合利用两个层级的标注进行训练，同时避免任一层发生灾难性遗忘。

这种分层训练不仅能生成捕捉长期信息（例如意图和时间依赖关系）的视频级表示，还能生成比传统单层方案学习到的更具表现力的片段级视频特征。这是通过我们的父子学习框架实现的：该框架要求聚合视频中的片段特征，使其与摘要捕捉到的长期上下文相匹配。

我们通过在时长达 3,670 小时的自我中心视角视频数据集 Ego4D 的叙述和摘要上进行训练，展示了模型的效果。 [13, 34]我们表明，HierVL 在多个视频基准测试中优于强大的基线方法和最先进方法，并成功迁移其预训练表征，用于在 Charades-Ego 上进行推理。 [74], EPIC-KITCHENS [16], 和 HowTo100M [63]。[1] 我们在两个层次上评估了这些表示。具体而言，在提交时，HierVL 在 Ego4D长期预测（LTA）、Charades-Ego动作识别和 EPIC-KITCHENS-100 上达到了最先进的性能多实例检索（零样本和微调设置）以及 HowTo100M 长视频分类。

[1]请注意，对于这些下游数据集和任务，我们不需要任何文本或摘要标注。

2. 相关工作

活动识别与检测。视频理解涵盖动作识别等任务 [24, 32, 48, 53, 86],动作预判 [2, 26, 28, 30, 60], 流程学习 [5, 8, 10, 64, 102], 以及动作定位 [90, 94, 101, 106]。各种视频数据集推动了这些方向的研究，包括 HowTo100M 等互联网视频集合[63], YouCookII [105], 和 CrossTask [106], , 以及新近录制的数据集，如 CharadesEgo [74], EPIC-KITCHENS [16], 和 Ego4D [13, 34]。作为训练资源，我们使用 Ego4D [13, 34], , 这是一个大规模、多样化的数据集，收录了世界各地日常活动的野外可穿戴摄像头视频。Ego4D 视频为摄像机佩戴者执行的每个动作都配有低层次文本描述（“叙述”），以及视频级摘要，因此非常适合我们的研究思路。

长视频表征。视频时长增加会带来计算瓶颈，使长视频理解变得困难。为使这项任务在计算上可行，人们提出了几种变通方案。传统方法包括使用预计算特征，以尽量减少反向传播的需求 [1, 20, 31, 85,95]，或降低帧率 [25, 38, 43, 46, 54, 87, 97,104, 107]。近期方法则通过构建“特征库” [84]或缓存记忆 [86]来缓解计算需求。结构化状态空间序列模型（S4）[35, 39] 将自注意力的二次复杂度降低为线性复杂度，从而能够高效训练长序列任务。另一种有望取得成效的方法是将细粒度片段级特征 [4, 27, 62, 67, 77, 78, 80, 82, 96] 聚合为整体视频表征，这通常用于视频分类任务。虽然这些方法都仅针对视频，我们提出了一种涵盖视觉和文本模态的多模态长视频表征。

视频与语言联合学习。 将视觉表示和语言表示投影到同一嵌入空间中的想法，已广泛用于多模态理解 [3, 55, 61, 63, 90]。这类联合表示支持多种任务，例如图像中的语言定位 [14, 21,、 57, 59, 72], 图像描述 [36, 51, 65, 81, 98], 和图像检索 [19, 37, 44, 49, 99], , 以及文本到视频检索 [11, 23, 58, 90, 103],、视频描述 [29, 58, 79, 89],和视频问答 [45, 47, 50, 69, 92, 93]。其中一些方法 [55, 58, 61, 63, 90] 采用对比学习（例如 InfoNCE [66]），以自监督方式将视频片段（或图像）与其叙述（或字幕）进行匹配。 [70] 中的自监督模型同时使用视觉和音频的窄窗口与宽窗口，并专注于短视频（例如 Kinetics 的 5 秒视频片段）。HERO [50] 在视频片段（时长几秒）与

其视频帧，仅使用片段级文本，而 [100] 则通过连接文本句子形成（非抽象式）段落来进行分层训练，从而增强父级理解能力，用于视频到段落检索和动作识别。

所有这些方法都只关注局部叙述/字幕。一个文本句子对应一个通常只有几秒钟的视频片段。选择较短时间窗口有两个原因：a）叙述通常只持续几秒钟；b）较长的视频片段会带来计算负担过重的问题，导致训练困难。相比之下，我们设计了一种分层方法，同时利用持续几秒钟的片段级叙述和持续数分钟的抽象式视频级摘要。我们表明，片段特征聚合使学习在计算上切实可行，而且采用这种分层文本描述能够提升片段级和视频级任务的表现。

3. 技术方法

我们提出了 HierVL，这是一种新颖的视频语言模型，能够捕捉片段级和视频级关系。图2概述了我们的方法。接下来，我们介绍注释（第3.1节），阐述嵌入学习方法（第3.2节），并讨论特征聚合策略（第3.3节）。最后，我们介绍损失函数（第3.4节）、训练过程（第3.5节）和实现细节（第3.6节）。

3.1. 分层视频标注

考虑一个经过分层标注的视频数据集，$\mathcal{D}_L = \{(V_i, N_i, S_i)\}_{i=1}^{|\mathcal{D}_L|}$ 其中 V_i 是一个长视频，N_i是描述视频中每个原子动作的一系列文本叙述，而 S_i 是对整个视频的高层文本摘要。用符号表示，$V_i = \{v_{ij}\}_{j=1}^{|V_i|}$ 是由一系列有序短片段组成的集合 v（每个片段时长为几秒），而 $N_i = \{n_{ij}\}_{j=1}^{|N_i|}$ 是由一系列有序叙述组成的集合n。请注意，我们不限制视频 V_i 的时间跨度，但在我们的实验中，视频通常以分钟计。例如，n_{ij} 可以是“他清洗画笔”或“他擦掉多余的颜料”，而高层摘要 S_i 则可以是“他在画室里作画”。片段 v_{ij} 包含对叙述n_{ij}的视觉演示，而 S_i 则是对完整视频V_i的抽象式摘要。其思路是让片段级表示捕捉视频中的细粒度动作，而视频级表示则应捕捉任务的总体目标。

我们利用 Ego4D 数据集 [13, 34] 来训练模型。Ego4D 包含 3,670 小时由全球 931 名不同的摄像机佩戴者拍摄的可穿戴摄像机视频，记录了日常生活活动。Ego4D 的标注中包括对摄像机佩戴者所执行每个动作的文本描述（“叙述”），以及视频级文本摘要，这些内容满足我们对 N 以及 S 的要求。自由形式的叙述由标注员在选定的时间点撰写，以记录所执行的每个动作。具体而言，标注员首先观看一段完整的 5 分钟视频，并为整体活动和环境撰写一则 1 至 3 句的简短摘要。随后，标注员被要求假装自己正在通过电话向一位看不到视频的朋友描述视频中发生的一切。由此得到时间上密集的逐步描述——平均每分钟 13.2 句，总计 385 万句（详情请参见 [34] 中的附录 D）。

3.2. 分层视频与文本联合嵌入

在我们的层次化设置中，我们有短期视频片段 v 和短期文本 n。我们希望学习短期表示 $f_v(v)$ 和 $f_n(n)$，分别称为视觉短期特征和文本短期特征。在长期层面，我们将多个 v 和多个 n分别汇集为 V 和 N。与此同时，我们希望学习长期表示$f_V(V)$ 和 $f_N(N)$（分别称为长期视觉特征和长期文本特征）。最后，我们有$f_n(S)$，即长期摘要特征；它通常由几句话组成，因此也使用 f_n进行编码。

目标是将 v，n，V，N，S投影到一个共同空间中，使语义相关的特征彼此接近。从数学上说，对于任意适当选取的相似度度量 $sim()$，以及$\forall i_1, i_2, j_1, j_2$，若 $(i_1, j_1) \neq (i_2, j_2)$，我们希望满足一项子级匹配约束：

$$sim\left(f_v(v_{i_1 j_1}), f_n(n_{i_1 j_1})\right) > sim(f_v(v_{i_1 j_1}), f_n(n_{i_1 j_2})) \tag{1}$$

以及 $\forall i, j$ 满足 $i \neq j$，以及 父级 匹配约束：

$$sim(f_V(V_i), f_n(S_i)) > sim(f_V(V_i), f_n(S_j)) \tag{2}$$

$$sim(f_N(N_i), f_n(S_i)) > sim(f_N(N_i), f_n(S_j)). \tag{3}$$

总的来说，式1意味着对应的短期表示应比不匹配的表示具有更高的相似度；式2 （以及式 3）意味着视频（及其叙述）与其摘要的相似度应高于与其他摘要的相似度。请注意，由于我们将短期和长期特征都投影到同一个空间中，因此如果语义相似，即使特征处于不同的层次级别，我们也允许它们在嵌入空间中彼此靠近。

3.3. 通过聚合高效获取长期特征

在视觉和文本模态中获取长期特征都颇具挑战。由于视频体量庞大，直接计算长期视觉特征需要更多资源，且往往会导致性能不佳和内存溢出 [4, 43, 84, 86]。自注意力模型是捕捉长期依赖关系的合适架构，但

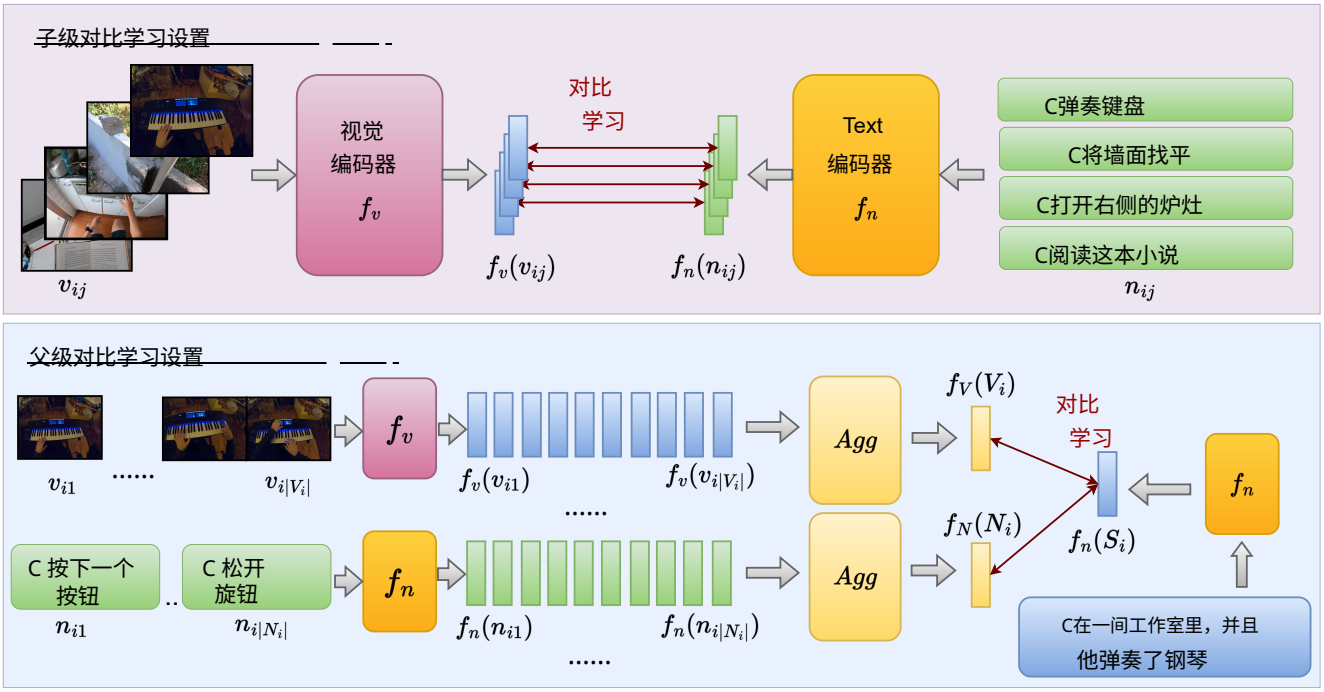


图2. 我们提出的方法示意图。在片段级对比学习设置（上方）中，我们将视频片段与其对应的叙述进行匹配。如图所示，一个批次中选取的片段来自不同的视频。在我们新颖的父级对比学习设置（下方）中，我们对短期特征进行采样，并将其聚合为长期表示，随后与摘要特征进行对比匹配。这些片段采样自同一个视频。请注意，f_v 和 f_n 在两个阶段中均为常见项，并且都可训练。（为简便起见，图中仅展示对比设置中的正样本对。）

由于 Transformer 模型对词元序列长度呈二次依赖，将其应用于大量文本句子（例如长文档）颇具挑战 [18]。Longformer [6] 通过多层级的全局和局部注意力缓解了这一问题。

受这些视觉和文本领域相关工作的启发，我们将短期特征聚合为长期表示 f_V 和 f_N。遵循这一策略，我们将长期视觉表示 f_V定义为 $f_V(V_i) = Agg\left(\{f_v(v_{ij})\}_{j=1}^{|V_i|}\right)$。类似地，长期文本表示 f_N 定义为 $f_N(N_i) = Agg\left(\{f_n(n_{ij})\}_{j=1}^{|N_i|}\right)$。我们考虑两种聚合函数 $Agg(.)$。第一种使用自注意力 Transformer 模块，以捕捉整个视频中的长期依赖关系。我们使用位置编码，使模型能够在视频级表示中嵌入时间顺序信息。我们将基于这一自注意力聚合器的模型变体称为 **HierVL-SA** 。我们考虑的第二种聚合方式是简单的平均池化（即无参数聚合器），它使所有短期特征对长期特征的贡献相同。该聚合器无法保留顺序信息。我们将这一版本命名为**HierVL-Avg**。我们在两种模态中使用相同的聚合器，

因为 $f(v)$和 $f(n)$具有相同的维度（事实上，在理想的对比训练中，对于匹配的视觉文本对，它们的取值相等）。

3.4. 对比预训练目标

如前所述，我们在两个层次上学习表示——子级 f_v, f_n 和父级 f_V， f_N。对于子级表示，预训练目标与先前工作类似 [55, 61, 63, 90]，后者将短期视觉表示与短期文本表示联系起来。具体而言，我们使用 EgoNCE 的一个变体 [55], ，这是一种兼顾动作和场景的 InfoNCE 变体 [66]。EgoNCE 将相似动作归为正样本，并将时间上接近的不同动作视为困难负样本。相比之下，我们省略后者，因为我们的层次化设置旨在 将具有相同相机佩戴者意图的不同动作聚合在一起。总体而言，短期预训练目标为:

$$\mathcal{L}_{child} = \frac{1}{|\tilde{\mathcal{B}}|} \sum_{i \in \tilde{\mathcal{B}}} \log \left(\frac{\sum_{j \in \tilde{\mathcal{P}}_i} \exp(f_v(v_i)^T f_n(n_j))}{\sum_{j \in \tilde{\mathcal{B}}} \exp(f_v(v_i)^T f_n(n_j))} \right)$$

~ ~其中 $\mathcal{B}$ 是短期特征的整体集合，而 $\mathcal{P}$ 是动作感知正样本的逐实例集合（详情请参阅 [55]。参见图2 （上）。

在父级，我们在 S-V 和 S-N 之间采用类似的预训练目标。参见图 2 （下）。如第3.3节所述，我们对 v 进行聚合，得到 V （并对 n 进行聚合，得到 N）。由于短期匹配已对比了 v 和 n，因此我们不会在父级再次对比 f_V 和 f_N。总体而言，长期预训练目标为 $\mathcal{L}_{parent} = \mathcal{L}^{SV}_{parent} + \mathcal{L}^{SN}_{parent}$ 其中

$$\mathcal{L}^{SV}_{parent} = \frac{1}{|\tilde{\mathcal{B}}|} \sum_{i \in \tilde{\mathcal{B}}} \log \left(\frac{\sum_{j \in \tilde{\mathcal{P}}_i} \exp(f_V(V_i)^T f_n(S_j))}{\sum_{j \in \tilde{\mathcal{B}}} \exp(f_V(V_i)^T f_n(S_j))} \right)$$

对于 $\mathcal{L}^{SN}_{parent}$，情况也类似。对于父级特征，摘要文本 S_i 的负样本既包括从 S_i的时间跨度之外选取的视觉表征，也包括文本表征。

3.5 训练策略

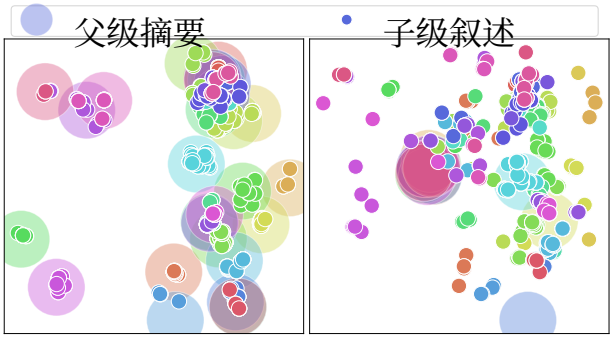
至此，我们讨论了用于层次视频-语言预训练的方法。为实现这一设置，我们采用联合训练方法。首先，我们训练 m 批短期视觉与文本配对数据（v, n），从而训练 f_v 和f_n。随后，我们训练一批长期特征，从而训练 f_V和 f_N。回顾一下，$f_V(.) = Agg(f_v(.))$，且 $f_N(.) = Agg(f_n(.))$。因此，在此批次中，我们会更新 Agg 的权重，以及短期 f_v 和 f_n。对比目标详见第 3.4 节。

同时训练两个层级的标注，其动机是确保函数 f_v 和 f_n 同时针对短期和长期特征进行优化，也就是说，两者都会受到文本摘要的影响。其他替代方案包括：(a) 分别为片段级和视频级特征使用不同模型，但这会增加模型参数，并使训练变得困难（无论是收敛方面还是 GPU 使用方面）；(b) 仅使用片段级数据训练，再针对视频级数据进行微调（或反过来），但众所周知，这类策略会导致灾难性遗忘 [33, 41, 42]。

图 3 展示了使用我们的方法学习到的 500 条摘要文本及其子级叙述的特征 f_n （左）以及 EgoVLP 的特征（右）。EgoVLP 中的摘要特征与叙述毫无关联，而 HierVL 则捕捉到了它们天然的层次结构，这一点可从嵌入空间中聚集在一起的相同颜色看出。特征的这种重塑体现了我们的片段级特征如何传达有关摄像机佩戴者更高层次意图的上下文信息。

3.6 实现细节

网络架构。 为了学习视频特征提取器 f_v，我们采用标准的 FrozenInTime [3] 视频骨干网络，它与 TimeSformer [7]略有不同，并受到视觉变换器的启发 [22]。基于 ViT 的视觉变换器常被用作特征提取器 [55, 68]，因为与其他骨干网络相比，它们的性能更出色。视频表示 f_v 从头开始学习；



HierVL-SA　　EgoVLP

图 3. 我们的 HierVL-SA（左）和 EgoVLP [55]（右）学习到的特征的 T-SNE 图。详情请参阅正文和补充材料。

输出表示是最终 CLS 词元的输出。对于短期片段，我们以 1 帧/秒的速度选取视频帧。接下来，文本特征提取器 f_n 采用 DistillBERT [71]架构，其性能与 BERT [18]相当，同时更加轻量。

聚合器。 我们的 HierVL-SA 变体采用 TimeSformer 架构 [7] 的 6 层自注意力模块实现，而 HierVL-Avg 则通过对特征取平均值实现。为保持批次大小恒定，对于 HierVL-SA 和 HierVL-Avg，我们从整个视频中均匀采样 16 个短期表示并进行聚合。

训练设置与参数。 我们使用 4 个节点对架构进行预训练，每个节点配备 8 块 32 GB NVIDIA V100 GPU，训练 10 轮次，耗时两天。我们使用 AdamW [56]优化器，学习率为 3×10^{-5}。每完成 $m = 5$ 轮次的片段级训练后，我们训练一个批次的视频级聚合。对于短期对比学习，每块 GPU 的批次大小为 16；对于长期视频级对比学习，每块 GPU 的批次大小为 1。请注意，一个视频级批次由同一视频的 16 个片段组成。

4. 实验

我们首先按照第3.6 节中讨论的设置和参数对我们的架构进行预训练，并在多个任务上报告结果，这些任务旨在直接评估所学习视频特征的质量（第4.1节）。接下来，我们展示预训练模型在涵盖短期和长期理解的各种下游任务上，提升了最先进水平（第4.2节）。

4.1. 预训练评估

我们使用 Ego4D [13, 34] 进行对比预训练。Ego4D 具有两级分层标注——短期逐步叙述，以及标注者观察到的演示的长期摘要。我们采用与 [55]相同的训练集和验证集划分。总体而言，共有3.8M 条短期叙述，以及 120K 条长期

方法	联合训练	Hier	Summ	聚合	摘要多项选择题	打乱多项选择题	EgoMCQ	
							视频间	视频内
EgoVLP [55]					—	—	90.6	57.2
EgoVLP（复现）	—	✗	✗	✗	89.0	20.0	90.1	**54.0**
HierVL-Avg（我们的方法）	✓	✓	✓	平均值	95.2	20.0	90.3	53.1
HierVL-SA（我们的方法）	✓	✓	✓	自注意力	**95.4**	**26.8**	90.5	52.4
HierVL-w/o Joint	✗	✗	✓	✗	89.8	24.2	72.0	29.4
HierVL-w/o Hier	✓	✗	✓	✗	93.7	20.0	90.7	50.5
HierVL-w/o Summ	✓	✓	✗	自注意力	20.0	22.1	**90.8**	52.1
HierVL-w/o Summ ↔ 叙述	✓	✓	✓	自注意力	94.7	26.1	90.4	50.0

表1. 在 Ego4D 预训练中，EgoMCQ、SummaryMCQ 和 ShuffleMCQ 上的预训练准确率，并与 EgoVLP（上方）和消融实验进行比较。对于所有验证集，随机水平对应 20.0 的准确率。我们提出的方法同时利用层次结构和长期摘要，在长期 SummaryMCQ 和 ShufleMCQ 任务上的表现优于所有基线方法。正如预期，两种方法在短期 MCQ 任务上的表现相当。—：不适用，粗体表示最佳，下划线表示次佳。

摘要标注。

预训练评估任务。 我们在 Ego4D 数据集上定义的三项任务中评估预训练质量：EgoMCQ（**多**项**选择**题，由 EgoVLP [55] 提出），以及我们提出的两个新基准测试——SummaryMCQ 和 ShuffleMCQ。在 **EgoMCQ** 中，模型会收到一条叙述提示和五个候选片段，并须将提示与正确的视频片段匹配，性能指标为准确率。视频内和视频间是验证数据的两个划分，候选视频片段分别选自同一视频或其他视频。**SummaryMCQ** 模仿 EgoMCQ 的视频-文本匹配测试，但在此任务中，模型收到的是一段**摘要** 和五个候选长期 视频选项。这些选项是覆盖整个摘要时长的视频。EgoMCQ 用于验证片段级性能，而 SummaryMCQ 用于验证视频级性能。最后，**ShuffleMCQ** 旨在评估时间理解能力：给定一段摘要文本，只有正确选项保留了片段间的时间顺序。其他四个视频选项则通过随机打乱原视频片段生成。

与 EgoVLP 的比较。我们主要与 EgoVLP [55], 进行比较，因为我们的模型采用了相同的架构，并在目标函数中使用其 EgoNCE 作为短期损失。然而，我们的方法利用层次化对比训练来使用摘要信息，而 EgoVLP 仅关注短期视觉-文本对应关系。对于 SummaryMCQ，我们采用无需参数的平均方法来计算聚合表示。

桌子1 展示了结果。[2] EgoVLP [55] 而我们的 HierVL 的两个变体在 EgoMCQ 上表现相近，这与该任务只需要短期信息这一事实一致。相比之下，HierVL-SA 在视频级（长期）任务 SummaryMCQ 和 ShuffleMCQ 上取得了更高的准确率。具体而言，HierVL-SA 在 SummaryMCQ 上的表现比 EgoVLP 高出 6% 以上。这凸显了我们的模型比 EgoVLP 聚合的短期特征更能有效捕捉长期意图。在 ShuffleMCQ 上，EgoVLP 和 HierVL-Avg 的表现都不优于随机水平（20%）。这反映出两种模型都无法捕捉时间顺序信息，而这对于区分原始摘要与打乱顺序的视频至关重要。相较之下，HierVL-SA 表现更出色，比这些模型高出 6.8%（相对提升 34%）。总之，我们的层次化学习在长期视频任务中表现突出，成功编码了事件之间的长期依赖关系。我们还观察到，在不同模型规模下，HierVL-SA 的表现均优于 EgoVLP。因此，进一步扩大模型规模也不会削弱我们架构的必要性（见补充材料）。

消融设计选择。 表 1的下半部分列出了 HierVL 的若干变体，用于对不同的设计选择进行消融。我们提出的架构包含三个不同的组件：(a) 一个在两个层级上运行的层次模型（父级摘要和子级叙述），(b) 使用文本摘要作为监督，以及 (c) 对这些分层标注进行联合训练。

HierVL-w/o Joint 是用于研究联合训练（组件 c）有效性的一个变体。我们使用 EgoVLP 的预训练权重初始化 HierVL-w/o Joint，并仅使用摘要训练整个网络（f_v, f_n, Agg），即不使用叙述。在此变体中，片段级表示通过父级损失得到间接监督。我们可以看到，尽管 HierVL-w/o Joint 在两个视频级任务上取得了不错的结果，但其在 EgoMCQ 上的性能远低于 EgoVLP 所达到的水平，而 EgoVLP 正是其初始化模型。这表明，仅凭摘要不足以监督强片段级表示的学习。

HierVL-w/o Hier 使用 (b, c)，而不使用 (a)，也就是说，我们使用

[2]第一行对应 EgoVLP [55]中报告的数值，第二行对应我们使用相同代码库复现的数值。我们将性能差异归因于硬件配置不同。



图4. 来自不同视频、代表不同短期步骤的视频片段，尽管它们在嵌入空间中彼此接近。两个视频具有相同的高层目标，即制作面包。

不依赖层次模型的摘要监督。我们将摘要文本标注随机分配给一个短期片段。值得注意的是，这一基线使用的监督量与我们提出的 HierVL 相同，但整体性能较低（EgoMCQ视频间上的表现略有提升）。这凸显了我们层次化训练方案的有效性。

HierVL-w/o Summ使用 (a, c)，但不使用 (b)，也就是说，监督信号并非来自摘要文本。请注意，这体现了 [100]的核心思想。用于对比学习的父级正样本是 f_V 和 f_N。这项消融实验旨在确定是否需要高层摘要，或者叙述聚合能否充当高层表示。我们观察到，在 SummaryMCQ 和 ShuffleMCQ 这两项视频级任务上，这一变体的效果明显不如 HierVL-SA。这是一项重要结果，因为它表明，HierVL-SA 有效地捕捉了人工标注者在摘要中表达的高层意图，而这种人工监督无法通过聚合短期叙述充分替代。

最后，**HierVL-w/oSumm↔Narr** 探讨了是否需要额外进行纯文本父级匹配，如式（3）所示。该消融实验考察了仅匹配 $f_V(V) \leftrightarrow f_n(S)$ 与同时匹配 $f_V(V) \leftrightarrow f_n(S)$和 $f_N(N) \leftrightarrow f_n(S)$的效果。我们发现，在子级和父级文本特征之间施加额外监督，确实能提升模型在所有验证集上的表现。

4.2. 下游评估

我们评估 HierVL 学得的表示在多个下游任务上的表现。

数据集。除了**Ego4D** [34]我们还使用**Charades-Ego** [74]，该数据集包含 7,860 个视频，视频分别从第一人称和第三人称视角录制，共涵盖 157 种动作

方法	动词编辑距离↓	名词编辑距离↓	动作编辑距离↓
Ego4D 基线 [34]	0.7389	0.7800	0.9432
Robovision [17]	0.7389	0.7688	0.9412
I-CVAE [60]	0.7526	0.7489	0.9308
HierVL-w/o Hier	0.7691	0.7454	0.9451
HierVL-Avg（我们的方法）	**0.7223**	0.7527	0.9401
HierVL-SA（我们的方法）	0.7239	**0.7349**	**0.9275**

表 2. Ego4D长期预测（LTA）挑战中的误差。ED 为 $Z = 20$处的编辑距离，越低越好。

类；**EPIC-Kitchens-100** [15, 16]，一个包含 100 小时自我中心视角视频的数据集，记录了在 4 个城市的 45 个家庭厨房中进行的非脚本化活动；以及 **HowTo100M** [63]，一个大型 YouTube 数据集，涵盖 23K 项视觉“操作指南”任务。

下游任务。 我们考虑以下任务：

- **长期动作预测（LTA）**。Ego4D 的 LTA 挑战要求模型根据当前动作（动词、名词）预测接下来的 20 个动作。评估指标为编辑距离（ED） [34]。

- **动作识别**。 Charades-Ego 的任务要求从 157 个类别中预测动作。评估指标为平均精度均值（mAP）。我们评估零样本和微调两种设置。

- **多实例检索（MIR）**。 EPIC-Kitchens-100 的 MIR 是一项文本到视频和视频到文本检索任务。对于 V→T 和 T→V，评估指标为平均精度均值（mAP）和归一化折损累积增益（nDCG）。我们报告二者的平均值。同样，我们在零样本和微调两种设置下进行评估。

- **视频分类**。 为展示我们预训练的迁移能力，我们在 HowTo100M 中出现频率最高的 100 个类别上进行线性探测。指标为分类准确率。

在全文中，我们报告了文献中现有最佳方法的相关比较结果，以及“无层次”消融实验；该实验使用了与 HierVL 完全相同的摘要数据/监督，从而精准揭示了我们分层训练理念的影响。

Ego4D LTA： 表 2 展示了 Ego4D LTA 挑战赛测试集上的结果。模型需要预测未来的 20 个动作，即使对人类而言也并非易事。我们在动词和名词预测方面均提升了最先进水平。此外，在提交时，我们的方法在公开排行榜上表现最佳（在表 2中，我们仅与已发表的工作进行比较）。尽管 HierVL-w/o Hier 同样能够访问摘要，但其表现并不理想，这进一步证明了我们分层训练的有效性。我们使用学习到的表示 f_v 和 Agg

零样本

方法	任务检查点 平均精度均值	预训练检查点 平均精度均值
EgoVLP [55]	25.0	19.4
HierVL-w/o Hier	24.6	24.5
HierVL-Avg（我们的方法）	25.2	23.9
HierVL-SA（我们的方法）	**26.0**	**25.0**

微调

方法	mAP
Actor [73]	20.0
SSDA [12]	23.1
I3D [12]	25.8
Ego-Exo [52]	30.1
EgoVLP [55]	32.1
HierVL-w/o Hier	32.6
HierVL-Avg（我们的方法）	32.6
HierVL-SA（我们的方法）	**33.8**

表 3. Charades-Ego动作识别的零样本（上）和微调（下）准确率。我们优于 EgoVLP，并且在零样本情况下抵御了过拟合。对于该数据集，我们的微调性能是迄今为止文献中报告的最佳结果。

零样本

方法	平均平均精度均值	平均归一化折损累计增益
EgoVLP [55]	16.6	23.1
HierVL-w/o Hier	17.8	24.1
HierVL-Avg（我们的）	16.7	23.5
HierVL-SA（我们的）	**18.9**	**24.7**

微调

方法	平均精度均值 平均值	平均归一化折损累计增益
MI-MM 使用 S3D [88]	29.2	44.7
MME [83] 使用 TBN [40]	38.5	48.5
JPoSE [83]使用 TBN [40]	44.0	53.5
EgoVLP [55]	45.0	59.4
HierVL-w/o Hier	44.7	59.8
HierVL-Avg（我们的方法）	44.9	59.8
HierVL-SA（我们的方法）	**46.7**	**61.1**

表 4. EPIC-KITCHENS-100 数据集上多实例检索任务的零样本和微调性能。

方法	推理	准确率
EgoVLP [55]	Avg	53.4
HierVL-SA（我们的）	自注意力	54.6
HierVL-Avg（我们的）	Avg	63.3
HierVL-SA（我们的）	Avg	**64.6**

表 5. HowTo100M 视频分类的线性探测结果。

后接一个多头解码器，如基线 [34] 中所示。该结果展示了我们学习到的特征聚合器（长期）以及短期视觉编码器的有效性 f_v。

Charades-Ego动作识别。表3（上）展示了零样本结果。EgoVLP [55] 报告称，从 Ego4D 迁移到 Charades-Ego 时出现了过拟合，因此选择了另一个预训练检查点。两个检查点的性能存在显著差距。我们报告了两种结果——表现最佳的预训练检查点（记为预训练检查点）以及 EgoVLP 选择的检查点（记为任务检查点）。我们的模型在迁移到 Charades-Ego 时不会过拟合；在相应检查点上的性能分别高出 5。6% 和 1。0%。在这项下游评估中，只需使用短期视觉编码器 f_v（冻结）。显然，我们的层次化预训练也改善了短期特征。

表 3（下半部分）展示了同一任务的微调结果。为了与最先进方法进行比较，我们从表现最佳的预训练检查点（HierVL-SA 的平均精度均值为 25.0% ）开始微调模型。我们的表现优于当前最先进的 EgoVLP [55]。我们针对该任务微调 f_v，结果显示短期特征有所改善。据我们所知，这是文献中针对该数据集报告的最佳结果。

EPIC-Kitchens-100 多实例检索。表4（上半部分）展示了零样本结果。我们观察到，最佳方法与我们的 HierVL-SA 之间相差2.3% 平均精度均值，且 1.6% 的提升。我们的 HierVL-Avg 也略优于最先进方法。在此任务中，我们同时使用短期编码器 f_v 和 f_n（二者均已冻结），因此该实验也验证了我们的主张：层次化学习能够改善短期表示。表4（下半部分）展示了同一任务的微调结果。我们对 f_v 和 f_n都进行了微调。与最先进方法相比，我们的两项指标均有所提升。

HowTo100M 视频分类。表5 展示了结果。在此线性探测设置中，所有 f_v、f_n 和 Agg均被冻结，只有一个额外的线性层可训练（可训练参数25.7K）。我们发现，我们所有学习到的表示都优于基线 EgoVLP。无参数平均在视频分类中效果良好 [4]。因此，我们加入了 HierVL-SA 的一个特殊变体：保留预训练的 f_v，并用平均操作替代 SA。这项额外实验也显示，与 HierVL-Avg 相比，HierVL-SA 中的短期特征 f_v 更具优势。

5. 结论

我们提出了一种新颖的层次化视频-语言嵌入。当前的嵌入对长期活动意图视而不见，而 HierVL 则同时关注短期的“这个人现在在做什么”和长期的“这个人打算做什么”。通过大量实验，我们证明这提升了短期和长期视频理解能力。我们的模型在多种视频挑战任务上推动了最先进水平，包括在 Charades-Ego 动作识别和 Ego4D 长期动作预测方面取得文献中的最佳整体表现。

致谢：感谢 Ziad Al-Halah 和 Tushar Nagarajan 对本文提出的意见。KG 是 Meta 的一名研究科学家。UTAustin 部分由 IFML NSF 人工智能研究所和 NSF-CCRI 提供支持。

参考文献

[1] Sami Abu-El-Haija、Nisarg Kothari、Joo nseok Lee、Paul Natsev、George Toderici、Balakrishnan Varadarajan 和 Sudheendra Vijayanarasimhan。Youtube-8m：一个大规模视频分类基准。arXiv 预印本 arXiv:1609.08675，2016。2[2] Yazan Abu Farha、Alexander Richard 和 Juergen Gall。你何时会做什么？——预测活动的时间发生事件。载于 IEEE计算机视觉与模式识别会议论文集，第 5343–5352 页，2018。2 – 。[3] Max Bain、Arsha Nagrani、G¨ul Varol 和 Andrew Zisser-man。Frozen in Time：用于端到端检索的视频与图像联合编码器。载于 IEEE/CVF国际计算机视觉会议论文集，第 1728–1738 页，2021。2、5[4] Max Bain、Arsha Nagrani、G¨ul Varol 和 Andrew Zisser-man。长视频检索搭便车指南。arXiv 预印本 arXiv:2205.08508，2022。2、3、8[5] Siddhant Bansal、Chetan Arora 和 CV Jawahar。我的视角才是最佳视角：从第一人称视频中进行流程学习。arXiv 预印本 arXiv:2207.10883，2022。2[6] Iz Beltagy、Matthew E Peters 和 Arman Cohan。Long-former：长文档 Transformer。arXiv 预印本 arXiv:2004.05150，2020。4[7] Gedas Bertasius、Heng Wang 和 Lorenzo Torresani。时空注意力是否就是视频理解所需的一切？载于 ICML，第 2 卷，第 4 页，2021。5[8] Jing Bi、Jiebo Luo 和 Chenliang Xu。通过上下文建模和基于模型的策略学习，在教学视频中进行程序规划。载于 IEEE/CVF国际计算机视觉会议论文集，第 15611–15620 页，2021。2[9] Fabian Caba Heilbron、Victor Escorcia、Bernard Ghanem 和 Juan Carlos Niebles。Activitynet：用于人类活动理解的大规模视频基准。载于 IEEE计算机视觉与模式识别会议论文集，第 961–970 页，2015。1[10] Chien-Yi Chang、De-An Huang、Danfei Xu、Ehsan Adeli、Li Fei-Fei 和 Juan Carlos Niebles。教学视频中的程序规划。载于欧洲计算机视觉会议论文集，第 334–350 页。施普林格，2020。2[11] Xing Cheng、Hezheng Lin、Xiangyu Wu、Fan Yang 和 Dong Shen。通过多流语料对齐和双重 Softmax损失改进视频文本检索。arXiv 预印本 arXiv:2109.04290，2021。2[12] Jinwoo Choi、Gaurav Sharma、Manmohan Chandraker 和 Jia-Bin Huang。无人机动作识别的无监督和半监督领域自适应。载于 IEEE/CVF冬季计算机视觉应用会议论文集，第 1717–1726 页，2020。8[13] Ego4D联盟。第一人称实时4D感知（Ego4D）数据库：一个大规模第一人称视频数据库，支持面向日常生活活动的多模态机器感知研究。https://sites.google.com/view/ego4d/home。2、3、5

[14] Bo Dai、Yuqi Zhang 和 Dahua Lin。利用深度关系网络检测视觉关系。载于《IEEE计算机视觉与模式识别会议论文集》，第 3076–3086 页，2017 年。2[15] Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Antonino Furnari、Jian Ma、Evangelos Kazakos、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 和 Michael Wray。重新审视自我中心视觉：EPIC-Kitchens-100 的采集、流程与挑战。《国际计算机视觉期刊》（IJCV），130 55，2022 33 7：–。[16] Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Antonino Furnari、Evangelos Kazakos、Jian Ma、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 等。重新审视自我中心视觉：EPIC-Kitchens-100 的采集、流程与挑战。《国际计算机视觉期刊》，130(1):33–55，2022 年。1, 2, 7[17] Srijan Das 和 Michael S Ryoo。视频+ CLIP 基线：用于 Ego4D 长期动作预判。arXiv 预印本 arXiv:2207.00579，2022 年。7[18] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本 arXiv:1810.04805，2018 年。4, 5[19] Haiwen Diao、Ying Zhang、Lin Ma 和 Huchuan Lu。用于图像文本匹配的相似性推理与过滤。载于《AAAI人工智能会议论文集》，第 35 卷，第 1218–1226 页，2021 年。2[20] Jeffrey Donahue、Lisa Anne Hendricks、Sergio Guadarrama、Marcus Rohrbach、Subhashini Venugopalan、Kate Saenko 和 Trevor Darrell。用于视觉识别与描述的长期循环卷积网络。载于《IEEE计算机视觉与模式识别会议论文集》，第 2625–2634 页，2015 年。2[21] Qi Dong、Zhuowen Tu、Haofu Liao、Yuting Zhang、Vijay Mahadevan 和 Stefano Soatto。利用部分与整体变换器及复合查询进行视觉关系检测。载于《IEEE/CVF国际计算机视觉会议论文集》，第 3550–3559 页，2021 年。2[22] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过 16x16 个词：用于大规模图像识别的变换器。arXiv 预印本 arXiv:2010.11929，2020 年。5[23] Han Fang、Pengfei Xiong、Luhui Xu 和 Yu Chen。CLIP2Video：通过图像 CLIP 掌握视频文本检索。arXiv 预印本 arXiv:2106.11097，2021 年。2[24] Christoph Feichtenhofer、Haoqi Fan、Jitendra Malik 和 Kaiming He。用于视频识别的SlowFast网络。载于《IEEE/CVF国际计算机视觉会议论文集》，第 6202–6211 页，2019 年。1, 2[25] Tsu-Jui Fu、Linjie Li、Zhe Gan、Kevin Lin、William Yang Wang、Lijuan Wang 和 Zicheng Liu。VIOLET：采用掩码视觉标记建模的端到端视频语言变换器。arXiv 预印本 arXiv:2111.12681，2021 年。2

[26] Antonino Furnari 和 Giovanni Maria Farinella。基于第一人称视频的动作预判滚动-展开 LSTM。IEEE模式分析与机器智能汇刊，43(11):4021–4036，2020 年。2[27] Adrien Gaidon、Zaid Harchaoui 和 Cordelia Schmid。利用动作原子进行动作时间定位。IEEE模式分析与机器智能汇刊，35(11):2782–2795，2013 年。2[28] Jiyang Gao、Zhenheng Yang 和 Ram Nevatia。RED：用于动作预判的强化编码器-解码器网络。arXiv 预印本 arXiv:1707.04818，2017 年。2[29] Simon Ging、Mohammadreza Zolfaghari、Hamed Pirsiavash 和 Thomas Brox。COOT：用于视频文本表示学习的协作式层次变换器。《神经信息处理系统进展》，33:22605–22618，2020 年。2[30] Rohit Girdhar 和 Kristen Grauman。预判式视频 Transformer。载于《IEEE/CVF国际计算机视觉会议论文集》，第 13505–13515 页，2021 年。2[31] Rohit Girdhar、Deva Ramanan、Abhinav Gupta、Josef Sivic 和 Bryan Russell。ActionVLAD：学习用于动作分类的时空聚合。载于《IEEE计算机视觉与模式识别会议论文集》，第 971–980 页，2017 年。2[32] Rohit Girdhar、Mannat Singh、Nikhila Ravi、Laurens van der Maaten、Armand Joulin 和 Ishan Misra。Omnivore：适用于多种视觉模态的单一模型。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 16102–16112 页，2022 年。2[33] Ian J Goodfellow、Mehdi Mirza、Da Xiao、Aaron Courville 和 Yoshua Bengio。对基于梯度的神经网络灾难性遗忘的实证研究。arXiv 预印本 arXiv:1312.6211，2013 年。5[34]Kristen Grauman、Andrew Westbury、Eugene Byrne、Zachary Chavis、Antonino Furnari、Rohit Girdhar、Jackson Hamburger、Hao Jiang、Miao Liu、Xingyu Liu 等。Ego4D：环游世界，记录 3,000 小时的自我中心视角视频。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 18995–19012 页，2022 年。1, 2, 3, 5, 7, 8[35] Albert Gu、Karan Goel 和 Christopher R´e。利用结构化状态空间高效建模长序列。arXiv 预印本 arXiv:2111.00396，2021 年。2[36] Xiaowei Hu、Zhe Gan、Jianfeng Wang、Zhengyuan Yang、Zicheng Liu、Yumao Lu 和 Lijuan Wang。扩展视觉-语言预训练以用于图像描述。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 17980–17989 页，2022 年。2[37] Yan Huang、Qi Wu、Chunfeng Song 和 Liang Wang。学习用于图像与句子匹配的语义概念及顺序。载于《IEEE计算机视觉与模式识别会议论文集》，第 6163–6171 页，2018 年。2[38] Noureldien Hussein、Efstratios Gavves 和 Arnold WM Smeulders。用于复杂动作识别的Timeception。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 254–263 页，2019 年。2[39] Md Mohaiminul Islam 和 Gedas Bertasius。使用状态空间视频模型进行长电影片段分类。arXiv 预印本 arXiv:2204.01692，2022。2[40] Evangelos Kazakos、Arsha Nagrani、Andrew Zisserman 和 Dima Damen。Epic-fusion：用于第一人称动作识别的视听时序绑定。载于《IEEE/CVF国际计算机视觉会议论文集》，第5492–5501页，2019。8[41] Ronald Kemker、Marc McClure、Angelina Abitino、Tyler Hayes 和 Christopher Kanan。衡量神经网络中的灾难性遗忘。载于《AAAI人工智能会议论文集》，第32卷，2018。5[42] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska 等。克服神经网络中的灾难性遗忘。《美国国家科学院院刊》，114(13):3521–3526，2017。5[43] Bruno Korbar、Du Tran 和 Lorenzo Torresani。Scsampler：从视频中采样显著片段以实现高效动作识别。载于《IEEE/CVF国际计算机视觉会议论文集》，第6232–6242页，2019。2, 3[44] Kuang-Huei Lee、Xi Chen、Gang Hua、Houdong Hu 和 Xiaodong He。用于图像文本匹配的堆叠交叉注意力。载于《欧洲计算机视觉会议论文集》（ECCV），第201–216页，2018。2[45] Jie Lei、Tamara L Berg 和 Mohit Bansal。揭示视频与语言学习中的单帧偏差。arXiv 预印本 arXiv:2206.03428，2022。2[46] Jie Lei、Linjie Li、Luowei Zhou、Zhe Gan、Tamara L Berg、Mohit Bansal 和 Jingjing Liu。少即是多：通过稀疏采样进行视频与语言学习的 Clipbert。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第7331–7341页，2021。2[47] Jie Lei、Licheng Yu、Tamara L Berg 和 Mohit Bansal。Tvqa+：用于视频问答的时空定位。arXiv 预印本 arXiv:1904.11574，2019。2[48] Kunchang Li、Yali Wang、Junhao Zhang、Peng Gao、Guanglu Song、Yu Liu、Hongsheng Li 和 Yu Qiao。Uniformer：统一卷积与自注意力以实现视觉识别。arXiv 预印本 arXiv:2201.09450，2022。2[49] Kunpeng Li、Yulun Zhang、Kai Li、Yuanyuan Li 和 Yun Fu。用于图像文本匹配的视觉语义推理。载于《IEEE/CVF国际计算机视觉会议论文集》，第4654–4662页，2019。2[50] Linjie Li、Yen-Chun Chen、Yu Cheng、Zhe Gan、Licheng Yu 和 Jingjing Liu。HERO：用于视频+ 语言全方位表征预训练的层次化编码器。arXiv 预印本 arXiv:2005.00200，2020。2[51] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei 等。Oscar：面向视觉语言任务的物体语义对齐预训练。载于《欧洲计算机视觉会议论文集》，第121–137页。施普林格，2020。2[52] Yanghao Li、Tushar Nagarajan、Bo Xiong 和 Kristen Grauman。Ego-exo：将视觉表征从第三人称视频迁移到第一人称视频

。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第6943–6953页，2021。8[53] Yanghao Li、Chao-Yuan Wu、Haoqi Fan、Karttikeya Mangalam、Bo Xiong、Jitendra Malik 和 Christoph Feichtenhofer。Mvitv2：用于分类和检测的改进型多尺度视觉变换器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4804–4814页，2022。1, 2[54] Ji Lin、Chuang Gan 和 Song Han。Tsm：用于高效视频理解的时间移位模块。载于《IEEE/CVF国际计算机视觉会议论文集》，第7083–7093页，2019。2[55] Kevin Qinghong Lin、Alex Jinpeng Wang、Mattia Soldan、Michael Wray、Rui Yan、Eric Zhongcong Xu、Difei Gao、Rongcheng Tu、Wenzhe Zhao、Weijie Kong 等。第一人称视频与语言预训练。载于 NeurIPS，2022。1, 2, 4, 5, 6, 8[56] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101，2017。5[57] Gen Luo、Yiyi Zhou、Xiaoshuai Sun、Liujuan Cao、Chenglin Wu、Cheng Deng 和 Rongrong Ji。用于联合指代表达理解与分割的多任务协作网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10034–10043页，2020。2[58] Huaishao Luo、Lei Ji、Botian Shi、Haoyang Huang、Nan Duan、Tianrui Li、Jason Li、Taroon Bharti 和 Ming Zhou。Univl：用于多模态理解与生成的统一视频与语言预训练模型。arXiv 预印本 arXiv:2002.06353，2020。1, 2[59] Junhua Mao、Jonathan Huang、Alexander Toshev、Oana Camburu、Alan L Yuille 和 Kevin Murphy。无歧义物体描述的生成与理解。载于《IEEE计算机视觉与模式识别会议论文集》，第 11 20, 2016 2 –。[60] Esteve Valls Mascaro 、 Hyemin Ahn 和 Dongheui Lee。Ego4D挑战赛2022：意图条件下的人类长期自我中心动作预测。arXiv 预印本 arXiv:2207.12080，2022。2, 7[61] Antoine Miech、Jean-Baptiste Alayrac、Lucas Smaira、Ivan Laptev、Josef Sivic 和 Andrew Zisserman。通过未经整理的教学视频端到端学习视觉表征。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9879–9889页，2020。1, 2, 4[62] Antoine Miech、Ivan Laptev 和 Josef Sivic。利用上下文门控进行可学习池化，以实现视频分类。arXiv 预印本 arXiv:1706.06905，2017。2[63] Antoine Miech、Dimitri Zhukov、Jean-Baptiste Alayrac、Makarand Tapaswi、Ivan Laptev 和 Josef Sivic。HowTo100M：通过观看一亿个旁白视频片段学习文本-视频嵌入。载于《IEEE/CVF国际计算机视觉会议论文集》，第2630–2640页，2019。1, 2, 4, 7[64] Zwe Naing 和 Ehsan Elhamifar。通过学习部分摘要完成流程。载于《英国机器视觉会议》，2020。2[65] Van-Quang Nguyen、Masanori Suganuma 和 Takayuki Okatani。Grit：利用双视觉特征打造更快、更出色的图像描述变换器。arXiv 预印本 arXiv:2207.09666，2022。2[66] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。通过对比预测编码进行表征学习。arXiv 预印本 arXiv:1807.03748，2018。2、4[67] Hamed Pirsiavash 和 Deva Ramanan。利用分段文法解析动作视频。发表于 IEEE计算机视觉与模式识别会议论文集，第 612–619 页，2014。2[68] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移视觉模型。发表于 Marina Meila 和 Tong Zhang（编），第 38 届国际机器学习会议论文集，机器学习研究论文集第 139 卷，第 8748–8763 页。PMLR，2021 年 7 月 18–24 日。5[69] Santhosh Ramakrishnan、Ziad Al-Halah 和 Kristen Grauman。NaQ：利用叙述作为查询来监督情景记忆。发表于 CVPR，2023。2[70] Adria Recasens、Pauline Luc、Jean-Baptiste Alayrac、Luyu Wang、Florian Strub、Corentin Tallec、Mateusz Malinowski、Viorica P˘atr˘aucean、Florent Altch´e、Michal Valko 等。拓宽视野，助力自监督视频学习。发表于 IEEE/CVF国际计算机视觉会议论文集，第 1255–1265 页，2021。2[71]Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。DistilBERT：BERT 的精简版本，更小、更快、更经济、更轻量。arXiv 预印本 arXiv:1910.01108，2019。5[72] Hengcan Shi、Hongliang Li、Fanman Meng 和 Qingbo Wu。用于指代表达图像分割的关键词感知网络。发表于欧洲计算机视觉会议（ECCV）论文集，第 38–54 页，2018。2[73] Gunnar A Sigurdsson、Abhinav Gupta、Cordelia Schmid、Ali Farhadi 和 Karteek Alahari。Actor 与观察者：联合建模第一人称和第三人称视频。发表于 IEEE计算机视觉与模式识别会议论文集，第 7396–7404 页，2018。8[74] Gunnar A Sigurdsson、Abhinav Gupta、Cordelia Schmid、Ali Farhadi 和 Karteek Alahari。Charades-Ego：一个大规模的配对第三人称和第一人称视频数据集。arXiv 预印本 arXiv:1804.09626，2018。1、2、7[75] Yansong Tang、Dajun Ding、Yongming Rao、Yu Zheng、Danyang Zhang、Lili Zhao、Jiwen Lu 和 Jie Zhou。COIN：用于全面教学视频分析的大规模数据集。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第 1207–1216 页，2019。1[76] Makarand Tapaswi、Yukun Zhu、Rainer Stiefelhagen、Antonio Torralba、Raquel Urtasun 和 Sanja Fidler。MovieQA：通过问答理解电影故事。发表于 IEEE计算机视觉与模式识别会议论文集，第 4631–4640 页，2016。1

[77] G¨ul Varol、Ivan Laptev 和 Cordelia Schmid。用于动作识别的长时域时间卷积。IEEE模式分析与机器智能汇刊，40(6):1510–1517，2017。2[78] Jue Wang、Gedas Bertasius、Du Tran 和 Lorenzo Torre-sani。视频变换器的长短时序对比学习。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 14010–14020 页，2022 年 6 月。2[79] Junke Wang、Dongdong Chen、Zuxuan Wu、Chong Luo、Luowei Zhou、Yucheng Zhao、Yujia Xie、Ce Liu、Yu-Gang Jiang 和 Lu Yuan。Omnivl：面向图像语言和视频语言任务的单一基础模型。arXiv 预印本 arXiv:2209.07526，2022。2[80] Jue Wang 和 Anoop Cherian。利用对抗扰动学习判别式视频表征。发表于欧洲计算机视觉会议（ECCV）论文集，第 685–701 页，2018。2[81] Jianfeng Wang、Zhengyuan Yang、Xiaowei Hu、Linjie Li、Kevin Lin、Zhe Gan、Zicheng Liu、Ce Liu 和 Lijuan Wang。Git：用于视觉和语言的生成式图像到文本变换器。arXiv 预印本 arXiv:2205.14100，2022。2[82] Limin Wang、Yuanjun Xiong、Zhe Wang、Yu Qiao、Dahua Lin、Xiaoou Tang 和 Luc Van Gool。时序分段网络：迈向良好的深度动作识别实践。发表于欧洲计算机视觉会议论文集，第 20–36 页。施普林格，2016。2[83] Michael Wray、Diane Larlus、Gabriela Csurka 和 Dima Damen。通过多种词性嵌入进行细粒度动作检索。发表于 IEEE/CVF国际计算机视觉会议论文集，第 450–459 页，2019。8[84] Chao-Yuan Wu、Christoph Feichtenhofer、Haoqi Fan、Kaiming He、Philipp Krahenbuhl 和 Ross Girshick。用于细粒度视频理解的长时特征库。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第 284–293 页，2019。2、3[85] Chao-Yuan Wu 和 Philipp Krahenbuhl。迈向长视频理解。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第 1884–1894 页，2021。2[86] Chao-Yuan Wu、Yanghao Li、Karttikeya Mangalam、Haoqi Fan、Bo Xiong、Jitendra Malik 和 Christoph Feichten-hofer。MemViT：用于高效长时视频识别的记忆增强多尺度视觉变换器。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第 13587–13597 页，2022。1、2、3[87] Chao-Yuan Wu、Manzil Zaheer、Hexiang Hu、R Manmatha、Alexander J Smola 和 Philipp Kr¨ahenb¨uhl。压缩视频动作识别。发表于 IEEE计算机视觉与模式识别会议论文集，第 6026–6035 页，2018。2[88] Saining Xie、Chen Sun、Jonathan Huang、Zhuowen Tu 和 Kevin Murphy。重新思考时空特征学习：视频分类中的速度-准确率权衡。发表于欧洲计算机视觉会议（ECCV）论文集，第305–321页，2018年。8[89]Hu Xu、Gargi Ghosh、Po-Yao Huang、Prahal Arora、Masoumeh Aminzadeh、Christoph Feichtenhofer、Florian Metze 和 Luke Zettlemoyer。VLM：面向视频理解的任务无关视频语言模型预训练。arXiv预印本 arXiv:2105.09996，2021年。2[90] Hu Xu、Gargi Ghosh、Po-Yao Huang、Dmytro Okhonko、Armen Aghajanyan、Florian Metze、Luke Zettlemoyer 和 Christoph Feichtenhofer。Videoclip：用于零样本视频文本理解的对比预训练。arXiv预印本 arXiv:2109.14084，2021年。1、2、4[91] Jun Xu、Tao Mei、Ting Yao 和 Yong Rui。MSR-VTT：用于连接视频与语言的大型视频描述数据集。载于 IEEE计算机视觉与模式识别会议论文集，第5288–5296页，2016年。1[92] Antoine Yang、Antoine Miech、Josef Sivic、Ivan Laptev 和 Cordelia Schmid。Just Ask：从数百万个带旁白的视频中学习回答问题。载于 IEEE/CVF国际计算机视觉会议论文集，第1686–1697页，2021年。2[93] Antoine Yang、Antoine Miech、Josef Sivic、Ivan Laptev 和 Cordelia Schmid。通过冻结的双向语言模型实现零样本视频问答。arXiv预印本 arXiv:2206.08155，2022年。2[94] Jianwei Yang、Yonatan Bisk 和 Jianfeng Gao。TACO：用于视频文本对齐的令牌感知级联对比学习。载于 IEEE/CVF国际计算机视觉会议论文集，第11562–11572页，2021年。2[95] Joe Yue-Hei Ng、Matthew Hausknecht、Sudheendra Vijayanarasimhan、Oriol Vinyals、Rajat Monga 和 George Toderici。Beyond Short Snippets：用于视频分类的深度网络。载于 IEEE计算机视觉与模式识别会议论文集，第4694–4702页，2015年 2 – 。[96] Joe Yue-Hei Ng、Matthew Hausknecht、Sudheendra Vijayanarasi mhan、Oriol Vinyals、Rajat Monga 和 George Toderici。Beyond Short Snippets：用于视频分类的深度网络。载于 IEEE计算机视觉与模式识别会议论文集，第4694–4702页，2015年 2 – 。[97] Rowan Zellers、Ximing Lu、Jack Hessel、Youngjae Yu、Jae Sung Park、Jize Cao、Ali Farhadi 和 Yejin Choi。Merlot：多模态神经脚本知识模型。神经信息处理系统进展，34：23634–23651，2021年。2[98] Yan Zeng、Xinsong Zhang 和 Hang Li。多粒度视觉语言预训练：将文本与视觉概念对齐。arXiv预印本 arXiv:2111.08276，2021年。2[99] Yan Zeng、Xinsong Zhang 和 Hang Li。多粒度视觉语言预训练：将文本与视觉概念对齐。arXiv预印本 arXiv:2111.08276，2021年。2[100] Bowen Zhang、Hexiang Hu 和 Fei Sha。视频与文本的跨模态层次建模。载于欧洲计算机视觉会议（ECCV）论文集，第374–390页，2018年。3、7

[101] Chenlin Zhang、Jianxin Wu 和 Yin Li。Actionformer：使用变换器定位动作时刻。*arXiv*预印本 *arXiv:2202.07925*，2022年。2[102] He Zhao、Isma Hadji、Nikita Dvornik、Konstantinos G Derpanis、Richard P Wildes 和 Allan D Jepson。P3iv：利用弱监督从教学视频中进行概率程序规划。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2938–2948页，2022年。2[103] Shuai Zhao、Linchao Zhu、Xiaohan Wang 和 Yi Yang。Centerclip：用于高效文本视频检索的令牌聚类。*arXiv*预印本 *arXiv:2205.00823*，2022年。2[104] Bolei Zhou、Alex Andonian、Aude Oliva 和 Antonio Torralba。视频中的时间关系推理。载于欧洲计算机视觉会议（*ECCV*）论文集，第803–818页，2018年。2[105] Luowei Zhou、Chenliang Xu 和 Jason J Corso。迈向从网络教学视频中自动学习程序。载于*AAAI*人工智能会议，第7590–7598页，2018年。2[106] Dimitri Zhukov、Jean-Baptiste Alayrac、Ramazan Gokberk Cinbis、David Fouhey、Ivan Laptev 和 Josef Sivic。基于教学视频的跨任务弱监督学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3537–3545页，2019年。1，2[107] Mohammadreza Zolfaghari、Kamaljeet Singh 和 Thomas Brox。ECO：用于在线视频理解的高效卷积网络。载于欧洲计算机视觉会议（*ECCV*）论文集，第695–712页，2018年。2