AnimalMotionCLIP: 在 CLIP 中嵌入运动以分析动物行为

Enmin Zhong Carlos R. del-Blanco Daniel Berjón Fernando Jaureguizar Narciso García
Grupo de Tratamiento de Imágenes (GTI), Information Processing and Telecommunications Center,
ETSI Telecomunicación, Universidad Politécnica de Madrid
{enmin.zhong, carlosrob.delblanco, daniel.berjon, fernando.jaureguizar, narciso.garcia}@upm.es

Abstract

最近，人们对将深度学习技术应用于动物行为识别产生了浓厚兴趣，特别是利用预训练的视觉语言模型（如 CLIP），因为它们在各种下游任务中具有显著的泛化能力。然而，将这些模型适应于特定领域的动物行为识别面临两个重要挑战：整合运动信息和设计有效的时序建模方案。本文提出 AnimalMotionCLIP 来解决这些问题，通过在 CLIP 框架内交错视频帧和光流信息来实现。此外，还提出了几种使用分类器聚合的时序建模方案并进行了比较：密集型、半密集型和稀疏型。结果表明，可以正确识别细粒度的时间动作，在动物行为分析中至关重要。在 Animal Kingdom 数据集上的实验显示，AnimalMotionCLIP 相较于最先进的方法具有更优的表现。

1. 介绍

理解动物行为对于野生动物保护、生态学和动物福利等领域至关重要。然而，自动识别动物行为由于不同物种之间以及同种动物之间的大小、形状和外观的多样性而面临重大挑战。此外，包括不同背景和栖息地在内的环境条件又增加了这一任务的复杂性。

当前自动动物行为识别方法主要依赖于视频深度学习技术 [1, 5, 6]，这些技术是在预定义的封闭行为集（类别）上进行训练的。这些行为通常局限于受控环境中的特定物种，例如实验室里的小鼠、畜牧业中的绵羊和奶牛 [3]，或者水产养殖业中的鲑鱼 [13]。然而，现实世界的情景中动物处于多样、复杂且杂乱的环境中，彼此互动或与周围环境交互。为应对这些挑战，通常需要大量的新标注数据来捕捉这些细微差别，因为难以在不同动物之间转移已学习的行为到新的行为类别。因此，这一策略成本高昂，并且无法随着广泛多样的动物物种进行扩展。现有的动物行为识别数据集非常稀缺，并且经常遇到与动物动作识别方法中类似的问题。例如 MammalNet[2] 和 KABR 数据集 [11] 受到其有限多样性的限制。具体来说，KABR 数据集包含三个动物类别的视频展示七种类型的行为，而 MammalNet 包括覆盖十二个行为类别的哺乳动物的视频。这种物种和行为范围受限的问题限制了有效训练和评估所需的数据多样性。大规模的 AnimalKingdom 数据集 [14] 解决了这些问题。它包含了 850 种哺乳动物、鱼类、两栖动物、爬行动物、鸟类和昆虫，具有跨越生命周期阶段、日常活动和社会互动（例如换毛、进食、玩耍）的 140 个动作类别。

最近，一种新的迁移学习范式出现了，基于像 CLIP（对比语言-图像预训练）[16] 这样的大型预训练视觉语言模型，这些模型能够根据成对的网络规模文本-图像数据集学习通用表示。它们在下游任务中表现出很强的泛化能力和可转移性，特别是在人类动作识别任务 [9, 15, 17, 21] 中。然而，与在人类动作识别中的应用相比，其在动物行为分析中的应用相对较少。目前将跨模态模型（文本-图像）应用于动物行为识别的研究往往采用类似于 ActionCLIP[20] 或 Vita-CLIP[21] 中用于人类动作识别的设计原则。这些方法通常使用基于提示的学习，其中为文本编码器输入设计了特定的提示向量，而保持整个模型冻结。尽管这种基于提示的方法增强了模型的表现，但结果往往难以解释并且容易受到嘈杂标签 [22] 的影响。例如，景等人 [8] 引入了一种

类别特定的提示技术，通过增加一个预测动物类别的分支来扩展 CLIP 模型。然而，这种动物类别专业化可能会偏向于特定的动物类别，并且会阻碍模型准确识别未见或代表性不足类别的行为的能力。蒙达尔等人[12] 提出了 MSQNet（多模态语义查询网络）用于人类和动物的一般动作识别。尽管融合视觉和语言信息有助于提高动物行为识别的性能，MSQNet 采用的方法需要在庞大的 Kinetics-400 数据集 [10] 上进行计算密集型的预训练，然后在目标 Animal Kingdom 数据集 [14] 上进行微调以完成预测任务。

最近提出的 WildCLIP[7] 结合了基于提示的技术和特征适配器来为野生动物图像生成注释。

与 WildCLIP 不同的是，我们的提案专注于视频级别的动物行为识别。

视频级别的动作识别涉及理解多个帧中运动的时态动态和序列，使其成为一个更复杂且计算密集型的任务。

先前工作的另一个常见局限是倾向于忽视运动信息和时间建模方案。这些挑战在动物领域尤为著名，因为细微的运动和时间信息可以导致不同的行为表现。关于运动信息，速度、加速度或移动模式的变化可以传达重要的行为线索，以区分不同的行为，如攻击性、探索或交配行为。光流可以从视频中获取所需的精细动态信息，捕捉到仅从静止帧中可能不易察觉的动物运动细微差别。另一方面，时间分辨率在区分不同行为方面也是基础性的，因为一个动作可能会发生在几秒、几分钟（甚至几小时）内。

为了解决之前的限制，提出了一种明确包含运动信息的视觉语言模型。此外，我们评估了不同的时间采样方案以识别最有效的方法。首先，在帧级别上将密集光流信息与颜色帧交错嵌入 CLIP 的预训练视觉-语言模型中，以嵌入精细的运动信息。然后，提出了几种时间分辨率模型并使用分类器组合进行比较。每个分类器关注的是根据密集、半密集和稀疏采样方案选择的不同帧集，以考虑一系列不同的全局行为上下文。因此，每个分类器产生一组部分推理得分，这些得分最终被汇总以预测视频中的行为。实验结果表明，我们的系统在动物行为识别方面具有显著的泛化能力，并且在 Animal Kingdom 数据集中比最先进的动作识别算法表现更好。

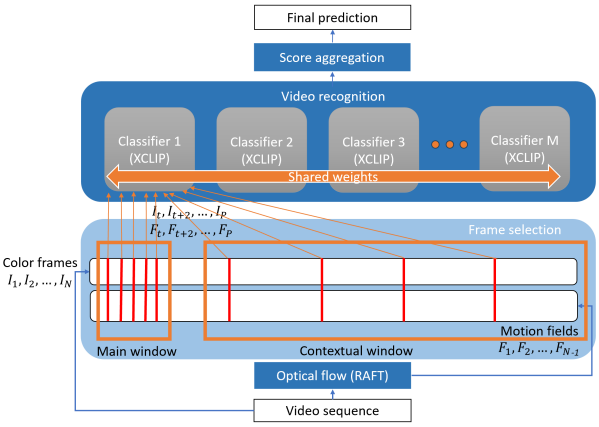

图 1. AnimalMotionCLIP 系统概述以及帧选择过程的示例说明。该系统处理视频序列中的彩色帧和运动场，使用主窗口和上下文窗口的组合。视频识别由多个共享权重的 XCLIP 分类器执行，随后进行分数聚合以生成最终预测。详情请参见 Sec. 2。

2. 提出的方法

所提出的 AnimalMotionCLIP 扩展并适应了视觉语言模型 CLIP 以进行视频动作识别。如 Fig. 1 所示，AnimalMotionCLIP 集成了三个主要模块：光流、视频识别和得分聚合。输入视频 $I = [I_1, I_2, ..., I_N]$ 经过光流模块处理，使用 [19]RAFT 算法获得相应的密集运动向量场 $F = [F_1, F_2, ..., F_{N-1}]$，其中 F_i 是在帧 I_i 和 I_{i+1} 之间计算的密集运动向量场。选择 RAFT 是因为它在近似运动边界和小位移精度方面的高准确性，这对于捕捉动物行为中的细微动作模式至关重要。然后，视频帧 I 和运动矢量场 F 由视频识别模块处理，该模块由一组基于 XCLIP[15] 的 M 分类器组成，这些分类器共享相同的权重但处理不同的视频时间内容。每个 XCLIP 分类器分析的视频内容依赖于两个窗口（Fig. 1 中的橙色矩形）：主要窗口和上下文窗口。一个分类器的主要窗口 W_m 专注于整个视频的一个特定小部分，使得其他分类器的主要窗口不与其重叠，并且所有主要窗口覆盖整个视频。提出了三种主要和上下文窗口的帧选择方案。对于密集方案，只从主要窗口均匀地抽取所有帧。对于半密集方案，从主窗口（由 Fig. 1 中的红线指示）均匀选取 $\frac{P}{2}$ 个视频帧。上下文窗口 W_c 覆盖了互补的 W_m 帧，从中提取其他 $\frac{P}{2}$ 个视频帧。最后，对于稀疏方案，不区分 W_m 和 W_c，而是从整个视频序列中随机准均匀选取帧，允许每个分类器的帧选择略有不

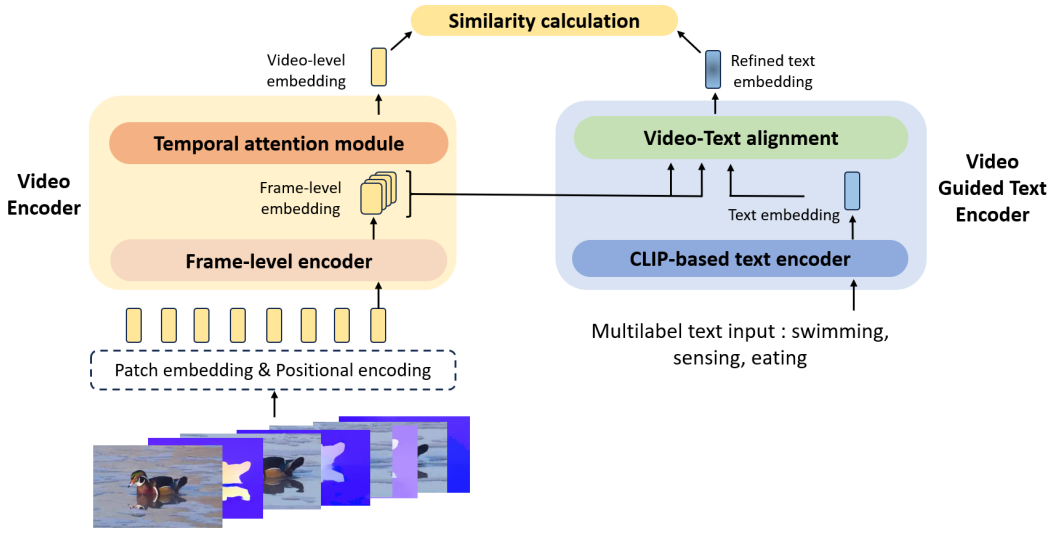

图 2. 视频识别模块中分类器架构的概述。它包含一个视频编码器（左）和一个视频引导文本编码器（右）。

同以获得更丰富的上下文。此外，对于每个采样方案，相应的运动矢量场都被绘制出来，并与视频帧交织作为 $IF = [I_{t_1}, F_{t_1}, I_{t_2}, F_{t_2}, ...]$。因此，每个 XCLIP 分类器处理不同的视频帧和运动矢量场。视频识别模块的输出是一组 M 评分向量，$S = [S_1, ..., S_M]$，每个分类器一个，表示对不同视频时态内容所考虑的动作/行为的概率。最后，聚合模块首先计算所有评分向量 S_M 的平均值，然后对其进行阈值处理以选择视频中概率最高的行为。请注意，该系统可以为每个视频预测多个行为。

视频识别模块 视频识别模块中的每个分类器均基于 XCLIP 架构，该架构已增强以包含视频和光流信息。该架构包括一个视频编码器和一个视频引导文本编码器。视频编码器通过结合颜色帧和运动向量场来计算包含整个视频及其底层运动的高语义表示的嵌入。视频引导文本编码器为描述行为的每个单词或文本计算一个嵌入。分类器使用对比学习 [16] 的概念进行训练，以生成具有相似嵌入的视频及其对应文本（指示一种或多种子行为），同时强制不相关文本和视频之间的差异性。为此，每个标签都使用了 sigmoid 激活函数和交叉熵损失函数来实现多标签（多类）分类。在推理过程中，应用 K-近邻策略通过将生成的视频嵌入与一组文本嵌入（每种描述感兴趣的潜在行为）进行比较来推断视频中的行为。

视频编码器可以进一步划分为三个子模块。第一个子模块，补丁嵌入和位置编码，将每个帧划分为 N 个大小为 $P \times P$ 的非重叠补丁，并将每个补丁映射到一个维度为 D 的嵌入中。对这些补丁嵌入应用位置编码以保留空间信息。随后，这一系列的补丁嵌入被传递给帧级编码器子模块，该子模块使用多头自注意力机制（MHSA）架构计算出一组图像嵌入，捕捉每个帧的高层次语义细节。这两个第一个子模块类似于基于视觉的转换器 [4] 的架构，但将概念扩展到处理多个帧而不是单个图像。最后，时间注意力子模块在时域上聚合这些帧级嵌入。使用标准的变换器架构，它对帧嵌入序列应用注意机制以生成一个单一的视频嵌入。这个最终的嵌入封装了视频中的内容和运动。

视频引导文本编码器由文本编码器和视频-文本对齐子模块组成。第一个组件利用预训练的基于 CLIP 的文本编码器来处理输入的多标签文本。每个单词经过词元嵌入和位置编码，以创建高度语义化的文本嵌入。接下来，视频-文本对齐子模块在视频级嵌入的指导下改进这些文本嵌入，增强两种模态之间语义相关性的捕捉。该策略替代了其他方法中用于文本编码器输入的手动设计固定提示模板，并为不同的视频内容和上下文提供了动态适应性。

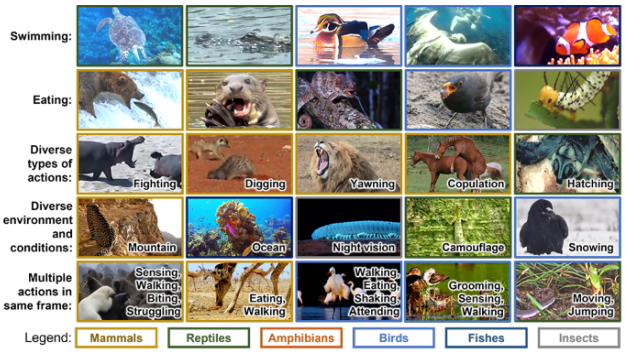

图 3. 动物王国数据集示例，包含不同环境中的多种动物物种及其多样行为。

3. 实验

为了评估所提出的 AnimalMotionCLIP 的有效性，我们在 Animal Kingdom 数据集 [14] 上进行了实验以解决动物行为一般识别的问题。遵循 Animal Kingdom 数据集的指南，使用平均精度均值（mAP）[18] 作为评估指标，其由公式给出

$$\text{mAP} = \frac{1}{n}\sum_{n}(R_n - R_{n-1})P_n, \tag{1}$$

其中 R_n 和 P_n 是在阈值 n 处的精度和召回率。

Animal Kingdom 数据集的作者 [14] 基于卷积神经网络展示了结果：I3D[1]、Slowfast[6] 和 X3D[5]。在这些模型中，X3D 达到了最高的平均精度均值（mAP），逐步将紧凑的 2D 图像分类架构扩展到 3D。然而，与我们提出的框架所实现的结果相比，其性能仍然相对较低（见表 1）。关于基于视觉语言模型的最新框架，有两项相关的工作与我们的工作密切相关。Category-CLIP[8] 和 MSQNet[12]。与这些方法相比，我们提出的 AnimalMotionCLIP 模型在动物行为分析方面显示出显著的进步（见表 1）。Category-CLIP 使用特定类别的提示技术增强了 CLIP 模型，通过增加一个额外的分支来预测动物类别，实现了 55.36%的平均精度均值（mAP）。同样地，MSQNet 利用视觉语言模型的泛化能力进行分类任务，达到了 55.59%的 mAP。相比之下，我们的 AnimalMotionCLIP 模型超越了这两种方法，并且在不依赖于额外的 Kinetic-400 数据集 [10]（该数据集专为人类动作识别设计）上计算成本高昂的预训练策略的情况下，实现了显著更高的 74.63% mAP。通过在这个数据集上进行预训练，MSQNet 学习了一组与动作相关的特征，然后将这些特征转移到动物行为识别中。甚至，在这种情况下，所提出的策略仍然超过了 MSQNet + 预训练的准确性。

方法	mAP(%)
I3D[1]	16.48
SlowFast[6]	20.46
X3D [5]	25.25
Category-CLIP[8]	55.36
MSQNet [12]	55.59
MSQNet + pre-training [12]	73.10
动物运动 CLIP（我们的）	74.63

表 1. 在 Animal Kingdom 数据集上的性能对比与最先进模型。最佳结果用粗体表示。

模态性	mAP(%)
Dense	68
Semi-dense	68.49
稀疏	73.93

表 2. 时间采样策略的性能比较。最佳结果用粗体表示。

消融研究 我们首先使用 RGB 模态评估了最有效的帧选择方法。如 Tab. 2 所示，密集、半密集和稀疏采样策略进行了比较，分别获得了 68%、68.49%和 73.93%的 mAP 分数。专注于详细片段的密集采样被认为不足。半密集采样显示出轻微改进，表明需要更具适应性的帧选择。使用准随机选择视频的稀疏采样方法证明是最有效的，更成功地捕捉到了多样化的动物行为和互动。随后，对于最有效的采样策略（稀疏的一种），我们进行了详细分析运动信息对行为识别性能的影响。如 Tab. 3 所示，仅使用 RGB 帧获得了 73.93%的 mAP，而仅使用光流帧得到了稍低的 65.1% mAP 分数。然而，结合 RGB 和光流帧导致了改进后的 74.63%mAP，展示了两种模态的互补性质，并强调了为准确的行为识别整合视觉和运动信息的重要性。

误差分析 在评估 AnimalMotionCLIP 用于动物行为识别时，我们观察到了几个影响所提框架性能的不准确来源。

模态	mAP(%)
RGB only	73.93
Optical flow only	65.1
RGB + **光流**	74.63

表 3. 不同模态之间的性能比较。最佳结果用粗体表示。

一个重要的问题来自于动作的重叠特征。例如，当真实标签是“跳跃”时，模型预测的是“保持静止”和“移动”的组合。这种错误可能是由于跳跃过程中的过渡性质造成的，在跳跃前后动物可能处于静止状态，导致模型区分这些片段而不是跳跃本身。此外，视频标注的准确性有时值得商榷，因为真实标签并不能完全代表所描绘的行为。例如，一段展示青蛙首先保持静止然后跳跃的视频仅被标记为“跳跃”，忽略了最初的静止阶段。另外，对于“感知”类，提出的框架有时会预测出如“注意”、“保持静止”和“感知”的动作组合。这表明模型可以识别相关的子行为，但难以将它们整合成一个单一且连贯的动作标签。另一个挑战是某些动作的复杂性，这些动作通常由多个子动作组成。例如，动物的梳理行为可能包括舔舐或抓挠自身等不同行动。我们的模型倾向于选择一系列动作来描述这一类别，对于单个梳理行为会产生如“注意”、“进食”、“保持静止”和“移动”的预测。这可能导致预测标签与真实标签之间存在差异。

总体而言，这些错误突显了需要更加精细的注释实践和增强模型能力，以便更好地捕捉和区分动物行为识别中的复杂和复合行为。

4. 结论

在本文中，我们介绍了用于动物行为分析的 AnimalMotionCLIP，将视觉语言模型适应于这一专业领域。我们的方法解决了两个关键挑战：整合运动信息和设计有效的时序建模方案。通过在 CLIP 框架内结合视频帧和光流，AnimalMotionCLIP 准确地关联了文本描述与观察到的动物行为。我们比较了各种时序建模方案，发现稀疏采样最有效地捕捉多样化的行为。利用预训练的视觉语言模型，我们的框架确保了效率并具有领域相关性，而无需大量预训练。在 Animal Kingdom 数据集上的实验表明，AnimalMotionCLIP 优于最先进的方法，在识别对动物行为分析至关重要的精细时序动作方面表现出色。

致 谢 此项研究由 TED2021-130225A-I00（ANEMONA）资助，该资助由 MCIN/AEI/10.13039/501100011033 提供，并由“欧盟下一代 EU/PRTR”提供支持。

这项工作还部分得到了西班牙政府的科学与创新部(AEI/FEDER)项目 PID2020-115132RB(SARAOS)的支持。

参考文献

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 1, 4

[2] Jun Chen, Ming Hu, Darren J. Coker, Michael L. Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding, 2023. 1

[3] Man Cheng, Hongbo Yuan, Qifan Wang, Zhenjiang Cai, Yueqin Liu, and Yingjie Zhang. Application of deep learning in sheep behaviors recognition and influence analysis of training data characteristics on the recognition effect. Computers and Electronics in Agriculture, 198:107010, 2022. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. 1, 4

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 1, 4

[7] Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. International Journal of Computer Vision, 2024. 2

[8] Yinuo Jing, Chunyu Wang, Ruxu Zhang, Kongming Liang, and Zhanyu Ma. Category-specific prompts

for animal action recognition with pretrained vision-language models. In Proceedings of the 31st ACM International Conference on Multimedia, page 5716 – 5724, New York, NY, USA, 2023. Association for Computing Machinery. 1, 4

[9] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding, 2022. 1

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 2, 4

[11] Maksim Kholiavchenko, Jenna Kline, Michelle Ramirez, Sam Stevens, Alec Sheets, Reshma Ramesh Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Eduardo Bessa, Isla Duporge, Tanya Berger-Wolf, Daniel I. Rubenstein, and Charles V. Stewart. Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), pages 31–40, 2024. 1

[12] Anindya Mondal, Sauradip Nag, Joaquin M Prada, Xiatian Zhu, and Anjan Dutta. Actor-agnostic multi-label action recognition with multi-modal query. In 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, 2023. 2, 4

[13] Håkon Måløy, Agnar Aamodt, and Ekrem Misimi. A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. Computers and Electronics in Agriculture, 167: 105087, 2019. 1

[14] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding, 2022. 1, 2, 4

[15] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022. 1, 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 3

[17] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners, 2023. 1

[18] Wanhua Su, Yan Yuan, and Mu Zhu. A relationship between the average precision and the area under the roc curve. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval, page 349 – 352, New York, NY, USA, 2015. Association for Computing Machinery. 4

[19] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 2

[20] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition, 2021. 1

[21] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting, 2023. 1

[22] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130 (9):2337 – 2348, 2022. 1