基于CLIP预训练模型的文本视频检索

魏晓静

摘要

摘要：文本视频检索在多模态研究中起着至关重要的作用，在日常生活中得到了广泛的应用，比如说抖音，bilibili 等视频软件中都有该项技术的应用。CLIP（Contrastive Language-Image Pre-training）模型是一种基于对比学习的多模态模型。在文本-视频检索中，目标是学习文本和视频之间的跨模态相似性函数，该函数将相关的文本-视频对排序高于不相关的文本-视频对。常见的与文本无关的聚合方案包括均值池或帧上的自关注，但这些很可能编码出在给定文本中未描述的误导性视觉信息。为了解决这个问题，本文使用 X-Pool 的跨模态注意力模型，该模型在文本和视频帧之间进行推理。在实验过程中使用 CLIP 预训练模型处理数据，通过跨模态注意力机制进行训练，在 MSRVTT 数据集上有不错的效果。

关键词：文本-视频检索、注意力机制、CLIP

1 引言

随着网络上每天上传的视频数量的增加，视频文本检索成为人们高效查找相关视频的一种新兴需求。除了实际的应用之外，视频文本检索是多模态视觉和语言理解的基础研究任务。

该论文介绍了一种名为 X-Pool 的新模型，该模型解决了文本-视频检索的挑战。文本-视频检索涉及在给定文本查询的情况下找到最具语义相关性的视频。挑战在于视频通常包含比文本可以捕捉的更广泛的信息，传统方法可能会聚合整个视频，而不考虑特定的文本，可能包括误导性的视觉信息。

X-Pool 采用了一种跨模态注意力模型，专注于与文本语义相似的视频子区域。它使用了一个经过缩放的点积注意力机制，允许文本关注视频中最相关的帧。这导致了一个基于文本对帧的注意力权重进行条件化的聚合视频表示。

该论文证明了 X-Pool 明显优于使用文本不可知视频池化的基线模型。它在 MSR-VTT、MSVD 和 LSMDC 等基准数据集上取得了新的最先进结果。这些发现强调并证明了基于文本描述提取视频中相关视觉线索的联合文本-视频推理的重要性。

2 相关工作

2.1 联合语言-图像理解

这是一种多模态学习形式，目的是理解并关联文本和图像模态。例如 CLIP、ALIGN、DeCLIP 和 ALBEF 等方法采用单模态编码器来学习一个联合潜在空间，该空间通过对比损失

匹配相关的文本-图像对。这些方法为后续的跨模态任务（如视觉问答、图像字幕和文本-图像检索）奠定了基础。

2.2 文本-视频检索

传统方法通常基于预训练的语言专家和视频专家的组合，然后通过后期融合整合语言和视觉流。例如 MoEE、CE、MMT、MDMMT 和 TeachText 等。这些工作的动机源于文本-视频检索使用的数据集规模较小。一些工作通过在大规模文本-视频数据集或通过文本-图像预训练来预训练自己的模型。然而，这些工作通常不允许直接推理给定文本与视频的最语义相似的子区域。最近的 CLIP4Clip 和 Straight-CLIP 等工作使用了在大规模文本-图像数据集上预训练的 CLIP 模型，但它们提出的视频聚合方案（包括均值池化、自注意力和多模态变换器）都不允许直接匹配文本与其最相关的视频子区域，这激发了本论文设计跨模态注意力模型的动机。

2.3 问题描述

文本-视频检索的目标是学习一个模型，以便在给定的文本 t 和视频 v 之间学习一个标量相似性函数 s(t,v)。目标是对相关的文本-视频对赋予较高的相似度，对不相关的对赋予较低的相似度。定义了两种检索任务：文本到视频的检索（t2v）和视频到文本的检索（v2t）。在 t2v 中，给定查询文本 t 和视频索引集 V，目标是根据它们与查询文本的相似度对所有视频 v 进行排名。类似地，在 v2t 中，给定查询视频 v 和文本索引集 T，目标是根据它们与查询视频的相似度对所有文本 t 进行排名。在这两个任务中，我们假设只有索引集在检索之前已知。问题的输入是一个视频 v 和一个文本 t。视频 v 被定义为一系列 F 个采样图像帧，即 v=[v1 v2 ⋯ vF]T，其中 vF 是分辨率为 H×W 的第 F 帧图像,文本 t 被定义为一系列分词化的词。

3 本文方法

在本节中，将逐步介绍激发最终模型 X-Pool 的见解和方法。首先会介绍怎么使用预训练的联合文本-图像模型，它是我们匹配文本和图像的模型的重要组成部分，我们将其扩展迁移到匹配文本和视频。之后解释了将视频聚合到与文本无关的嵌入中的缺点，并在提出了一个替代框架，该框架根据给定的文本聚合帧。然后，介绍了论文中的 X-Pool 模型，这是一个跨模态注意力模型，可以在文本和视频帧之间进行联合推理。我们的模型学习使用与给定文本语义最相似的帧来聚合视频。在最后详细描述了具体的优化操作。

3.1 扩展联合文本-图像模型

在已有的论文中证明了联合预训练的文本-图像模型在进行匹配语义相似的文本和图像任务中的能力。我们可以利用这些模型的现有文本-图像推理来引导联合文本-视频模型。预训练的联合文本-图像模型的丰富跨模态理解可以实现使用更少的视频数据学习语言-视频交

互，并在训练期间提供更高效的计算解决方案。

我们从 CLIP 开始，是因为它具有强大的下游性能，它的使用简单性，并且能够更客观地与最新的使用 CLIP 的方法进行对比，尽管其他预训练的文本图像模型可能是更合适的 backbone。为了从 CLIP 中引导文本视频检索，我们首先将文本和单个视频帧嵌入到其联合潜在空间中，然后将帧嵌入合并以获得视频嵌入。由于从预训练的 CLIP 模型中提取的现有信息包含丰富的文本-图像语义，我们使用 CLIP 作为主干来学习新的联合潜在空间来匹配文本和视频。

3.2 文本无关模型缺点及替代框架

在大多数现有的工作中，聚合函数 ρ 不直接考虑输入文本，而纯粹是视频帧的函数，例如 CLIP4clip 中提出的通过均值池化、自注意力机制等。

虽然将时间聚合函数定义为与文本无关是一个简单的基准，但这种方法存在一些重要的缺点。视频本质上比文本更具表现力，因此文本中捕获的信息通常不能完全捕获整个视频的信息。相反，文本在语义上与视频的某些子区域最为相似，我们将其定义为帧的子集，如图 1 所示。

因此，将整个视频聚合在一起的通用文本不可知聚合方案可能会编码输入文本中未描述的虚假信息。我们希望检索模型通过将其注意力集中在给定文本中描述的最相关的视频子区域上，可以对内容具有多样性的视频具有一定的鲁棒性。

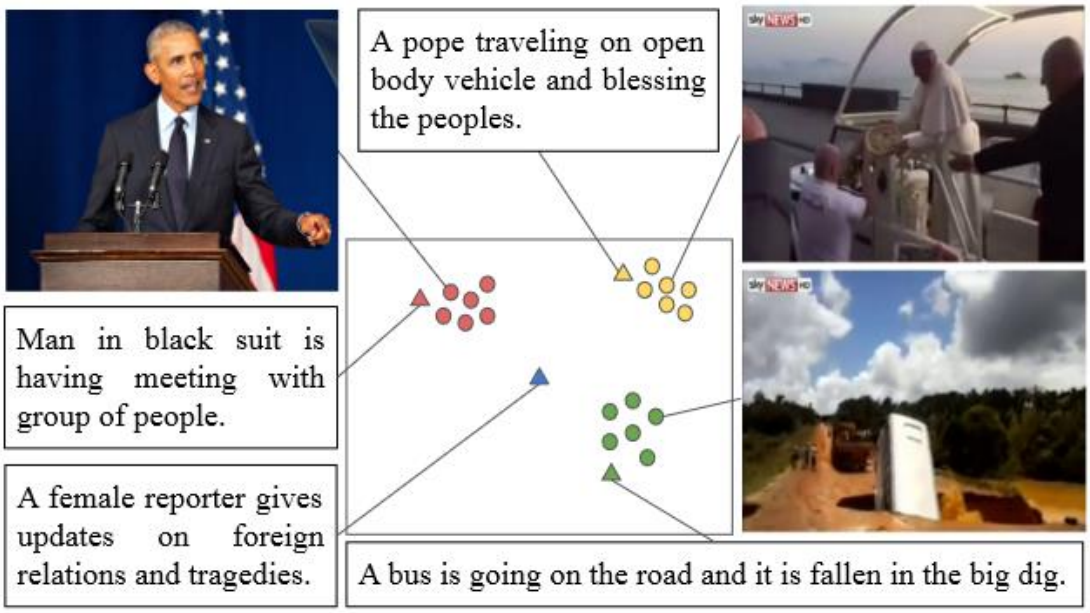

图 1：从 MSRVTT 数据集中逐字提取的单个视频及其标题的联合文本和视觉表示的说明。由于视频捕获的内容比单个文本更多，因此不管输入文本如何聚合整个视频可能会产生误导。

3.3 本文具体模型

我们注意到，重要的不是将文本与视频的整个内容匹配，而是与那些在语义上与给定文

本最相似的视频帧匹配。根据给定的文本，语义上最相似的帧可能会有所不同，因此可能存在多个相同有效的文本来匹配特定的视频，因此，我们的时间聚合函数应该直接在给定文本和视频帧之间进行推理。为此，我们定义了一个新的时间聚合函数 π，它允许我们聚合语义上与给定文本 t 最相似的视频帧。通过将 π 作用于 t，我们可以从视频 v 中提取 t 中描述的最相关的信息，同时抑制噪声和误导性的视觉线索。我们将得到的聚合视频嵌入记为 zv|t，并定义相似度函数 s(t, v)为:

$$\mathbf{z}_{v|t} = \pi(C_v \mid t)$$

$$s(t, v) = \frac{\mathbf{z}_t \cdot \mathbf{z}_{v|t}}{\|\mathbf{z}_t\| \|\mathbf{z}_{v|t}\|}$$

提出了一种结合文本条件池化的参数化方法来解决这些额外的考虑。名为 X-Pool 的跨模态注意力机制。

关于该模型的具体信息如图 2：

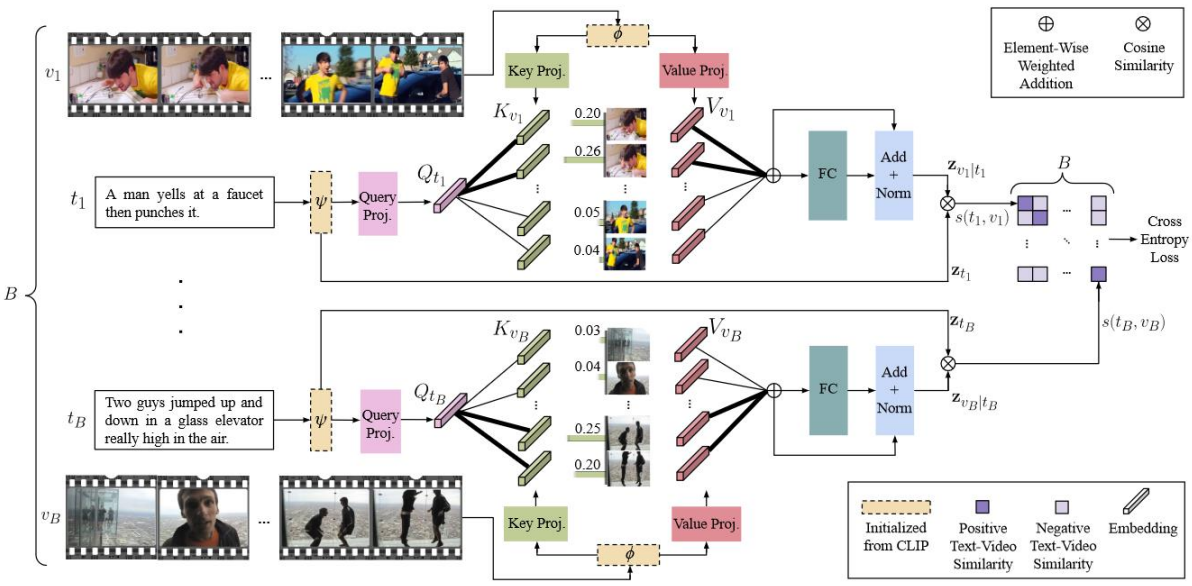

图 2：X-Pool 模型。对于给定的文本 t1，我们将其嵌入文本编码器 ψ，然后应用查询投影得到 Qt1。同样地，我们用图像编码器 φ 嵌入给定视频 v1 的帧，然后应用键投影来获得 Kv1。我们计算它们之间的注意力点积，如图中间的水平线图所示。我们的注意力机制允许 X-Pool 专注于给定输入文本的最相关框架。我们聚合了一组单独的值投影帧嵌入，我们通过先前计算的点积注意力分数来加权，以获得聚合的视频嵌入，然后我们通过具有残差连接的完全连接层(FC)来获得 zv1|t1。我们计算相似分数 s(t1, v1)作为 zv1|t1 和 zt1 = ψ(t1)之间的余弦相似度。最后，我们在获得 s(ti, vj)后计算交叉熵损失，正如刚才描述的那样，用于批量大小为 B 的每对(ti, vj)。

文章中的想法是设计一个具有参数能力的学习帧聚合函数，用于对视频中文本中语义最相似的帧进行跨模态推理，称之为 X-Pool。核心机制是该模型对文本和视频帧之间的缩放点积注意力的适应。在这些帧的条件下，我们生成一个视频嵌入，学习捕获给定文本中描述的语义上最相似的视频子区域。由于具有最高语义相似性的框架可能因文本而异，我们的缩放点积注意机制可以学习突出显示与给定文本相关的框架，同时抑制文本中未描述的框架。我

们的模型基于与给定文本的相关性有选择性地选择框架的能力是由与前面描述的 top-k 方法中概述的相同的文本条件反射见解所激发的。然而，与 top-k 方法不同的是，我们提出的模型学习了文本-视频对提取的最佳信息量，从而消除了手动指定 k 值的需要。此外，我们的交叉注意模块处理高相关性和低相关性框架，而不是像 top-k 方法那样采用硬选择相关框架。

本文中损失函数使用交叉熵损失函数。使用由 N 个文本和视频对{(ti, vi)}组成的数据集 D 训练模型。N i=1。在每对中，文本 ti 是对应视频 vi 的匹配文本描述。我们采用[44]中的交叉熵损失，将匹配的文本-视频对视为正，并将批处理中所有其他成对的文本-视频组合视为负。具体来说，我们共同最小化对称文本到视频和视频到文本的损失，具体损失函数公式定义如下：

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s(t_i,v_i)\cdot\lambda}}{\sum_{j=1}^{B}e^{s(t_i,v_j)\cdot\lambda}}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s(t_i,v_i)\cdot\lambda}}{\sum_{j=1}^{B}e^{s(t_j,v_i)\cdot\lambda}}$$

$$\mathcal{L} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t}$$

其中 s(ti, vj)是文本 ti 和视频 vj 之间的余弦相似度，B 是批大小，λ 是一个可学习的缩放参数。通过预先训练的 CLIP 模型和我们的跨模态注意机制，使用这种损失进行训练使我们的模型能够学习将文本与其 ground-truth 视频中语义最相似的子区域进行匹配。

4 复现细节

4.1 与已有开源代码对比

本工作使用 X-Pool 模型提供的源码，通过 CLIP 预训练模型处理数据，完成了 MSR-VTT 数据集中的文本-视频检索任务。

4.2 实验环境搭建

代码运行环境：Linux-18.04-Ubuntu；Python 版本：3.8；相关依赖：PyTorch 1.8.1，Transformers 4.6.1，OpenCV 4.5.3；CLIP: clip-vit-base-patch32

5 实验结果分析

5.1 实验数据集

MSR-VTT 由 10,000 个视频组成，每个视频都配有大约 20 个人工标记的字幕。我们注意到，MSR-VTT 中每个视频的多个字幕通常描述不同的视频子区域，这符合我们将给定文本与

视频中最相关的帧进行匹配的动机。该数据集中的视频长度从 10 秒到 32 秒不等，文章使用了 9k-Train 训练分割，以有效地与之前的作品进行比较，9k-train 由中拆分后的大约 9k 个视频组成。为了评估模型，我们使用了 1KA 测试集，该测试集由 1,000 个选定的字幕视频对组成。

5.2 定量分析结果

关于定性结果分析，做了可视化实验，展示 x-pool 模型的注意力机制。具体结果如下图 3 所示：

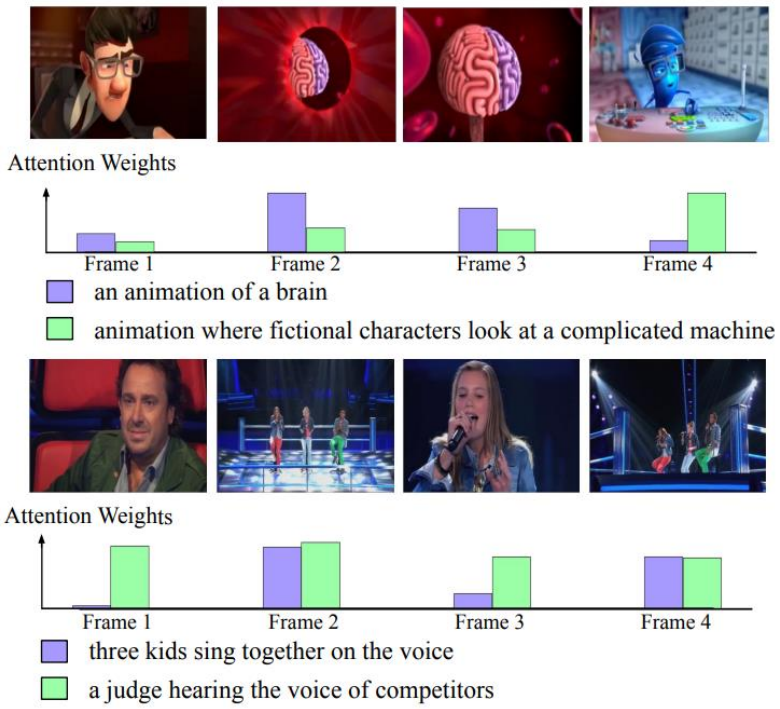

图 3：对于上面显示的每个帧，条形图显示了给定特定文本的模型中的注意力权重。

对于以上示例，显示了来自视频的四个采样帧，以及表示 X-Pool 从给定文本到每个帧的相关注意力权重的条形图。在上面的例子中，我们可以看到，当输入文本描述大脑动画时，我们的模型在中间帧输出更高的注意力权重，而在其他地方输出更低的注意力权重。另一方面，当输入文本描述的是一个正在看机器的虚构角色时，注意力权重就会相应地激活文本最相关的最后一帧。中间的第二个例子是歌唱比赛。在这里，“a judge hearing the voice of competitors”的文本描述了一个需要对所有框架进行推理的项目。通过以上可视化结果可以发现，X-Pool 关注整个视频，说明了该方法具有灵活性。

5.3 实验结果：

为了评估提出的方法，我们将其性能与最近的文献作品进行比较。在 MSR-VTT 数据集中以 9k 视频为训练集，1k 视频为测试集，得到图 4 实验结果。其中评价指标 R@k 指的是 top-k 召回中正确结果的比例；MdR 指的是正确结果在排序中的中位数；MnR 指的是正确结果在排序中的平均数。

	MSRVTT-9K				
Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE	20.9	48.8	62.4	6	28.2
MMT	26.6	57.1	69.6	4	24
Straight-CLIP	31.2	53.7	64.2	4	-
Support Set	30.1	58.5	69.3	3	-
MDMMT	38.9	69	79.7	**2**	16.5
Frozen	31	59.5	70.5	3	-
TeachText-CE+	29.6	61.6	74.2	3	-
CLIP4Clip-meanP	43.1	70.4	80.8	**2**	16.2
CLIP4Clip-seqTransf	44.5	71.4	81.6	**2**	15.3
X-Pool	**46.9**	**72.8**	**82.2**	**2**	**14.3**

图 4：X-Pool 模型实验结果

通过以上结果可以看出 X-Pool 模型比现有的工作有更好的效果。

由于原论文中的实验结果对实验设备要求略高，故根据原文中代码，在条件允许的范围内，调整相关超参数，将 batchsize 降为 8，其他相关超参数微调，使得 R@10，MdR 不变，MnR 结果得到了提升；R@1、R@5 略有下降。具体结果如下图 5：

	MSRVTT-9K				
Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
X-Pool	**46.9**	**72.8**	**82.2**	**2**	**14.3**
	44.2	71.4	82.2	**2**	**13.225**

图 5：原文超参数微调后实验结果。

6 总结与展望

在这篇论文中，强调了文本不可知视频池的缺点，并提出了用于文本视频检索的文本条件池的替代框架。然后为文本和视频帧之间的跨模态注意力设计了一个参数化模型，称为X-Pool。展示了X-Pool如何学习关注给定文本中最相关的帧，这也使该模型对视频内容多样性(例如场景转换的形式)具有更强的鲁棒性。在未来的工作中，我将通过阅读文献继续尝试不同的改进该模型方法，比如说加入监督机制、强化学习算法使更相似的样本获得更高的分数等。由于文章发表于2022年，在后续还会选择更SOTA的文章进行实现。

参考文献

[1] Satya Krishna Gorti, Noel Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-Pool: Cross-Modal Language Video Attention for Text-Video Retrieval. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 4996–5005

[2] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860,

2021

[3] Chen Jiang, Hong Liu, Xuzheng Yu, Qing Wang, Yuan Cheng, Jia Xu, Zhongyi Liu, Qingpei Guo, Wei Chu, Ming Yang, Yuan Qi: Dual-Modal Attention-Enhanced Text-Video Retrieval with Triplet Partial Margin Contrastive Learning. arXiv preprint arXiv:2309.11082v2

[4] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. NeurIPS, 2(6):7, 2020.

[5] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. arXiv preprint arXiv:2003.03186, 8, 2020.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425– 2433, 2015.

[7] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisser- ¨ man. Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv preprint arXiv:2104.00650, 2021.

[8] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe ˇ Morency. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2):423–443, 2018.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[10] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pages 190–200, 2011.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In European conference on computer vision, pages 104–120. Springer, 2020.

[12] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11583–11593, 2021.

[13] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3354–3363, 2021.

[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK,

August 23–28, 2020, Proceedings, Part IV 16, pages 214–229. Springer, 2020.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918, 2021.

[16] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7331–7341, 2021.

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 34, 2021.

[18] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137. Springer, 2020