arXiv:2601.12768v1 [cs.CV] 19 Jan 2026

深入探索：面向稳健视频文本检索的分层视觉感知

谢泽群*† 张博云*† 林宇骁† 金涛†

* 这些作者对本工作的贡献相同。
† 浙江大学 通讯作者为金涛。

摘要

视频文本检索（VTR）旨在根据自然语言查询定位相关视频。当前方法通常基于 CLIP 等预训练模型，但受到视频固有冗余以及依赖粗粒度最终层特征的限制，匹配准确率因此受限。为解决这一问题，我们提出 **HVP-Net（分层视觉感知网络）**，这是一种通过从视觉编码器的多个中间层提取并精炼特征来挖掘更丰富视频语义的框架。我们的方法在不同语义层次上，逐步从原始图像块令牌中提炼显著视觉概念，在减少冗余的同时保留对齐所需的关键细节。由此得到的视频表示更加稳健，并在 MSRVTT、DiDeMo 和 ActivityNet 等具有挑战性的基准上取得了新的最先进性能。我们的工作验证了利用层次化特征推动视频文本检索发展的有效性。代码见 https://github.com/boyun-zhang/HVP-Net。

索引词— 视频文本检索、视觉 Transformer、跨模态对齐、视频表征学习、计算机视觉

1. 引言

跨模态检索旨在将不同数据模态投射到共同的潜在空间中进行相似度度量，从而弥合它们之间固有的异质性差距 [1, 2]。这一范式及先进的特征融合技术已成功应用于多个领域，包括基于文本的人物检索 [3, 4], 金融风险分析 [5, 6],虚假新闻检测 [7, 8], 和序列推荐[9, 10]。在视频文本检索（VTR）中，通常通过利用 CLIP [11] 等大型预训练模型来学习联合的视频文本嵌入空间。然而，一个关键局限依然存在：这些方法，即使采用了更先进的多粒度匹配策略 [12, 13, 14],，也主要依赖视觉编码器最后一层的特征 [15, 16, 17]。这种做法不加区分地编码所有视觉信息，导致了严重的信息冗余

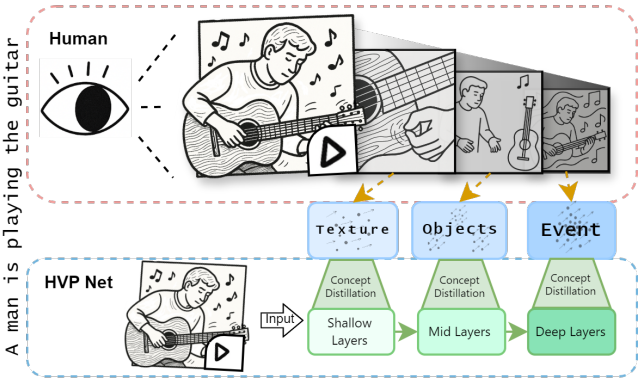

图 1. HVP-Net 的概念示意图，其设计灵感源自人类的层次感知过程。我们的模型通过浅层、中层和深层处理视觉数据，提取从低层纹理到高层事件的各类概念，以此模拟这一过程。

标记。更重要的是，这种做法忽略了编码器中间层所蕴含的丰富层次语义和动态信息 [18], ，从而限制了表征效率和匹配精度。

2. 方法

本节详细介绍我们提出的**分层视觉感知网络（HVP-Net）**。如图 2 所示，我们的核心思想是从预训练视觉编码器的多个中间层提取并优化特征，从而构建鲁棒的视频表征。该过程包含三个主要阶段：层次特征编码、一种新颖的图块处理模块，以及通过对比学习优化的多粒度对齐。

2.1. 层次特征编码

我们使用预训练的 CLIP 模型 [11] 作为骨干网络。从其文本编码器中，我们提取全局句子特征

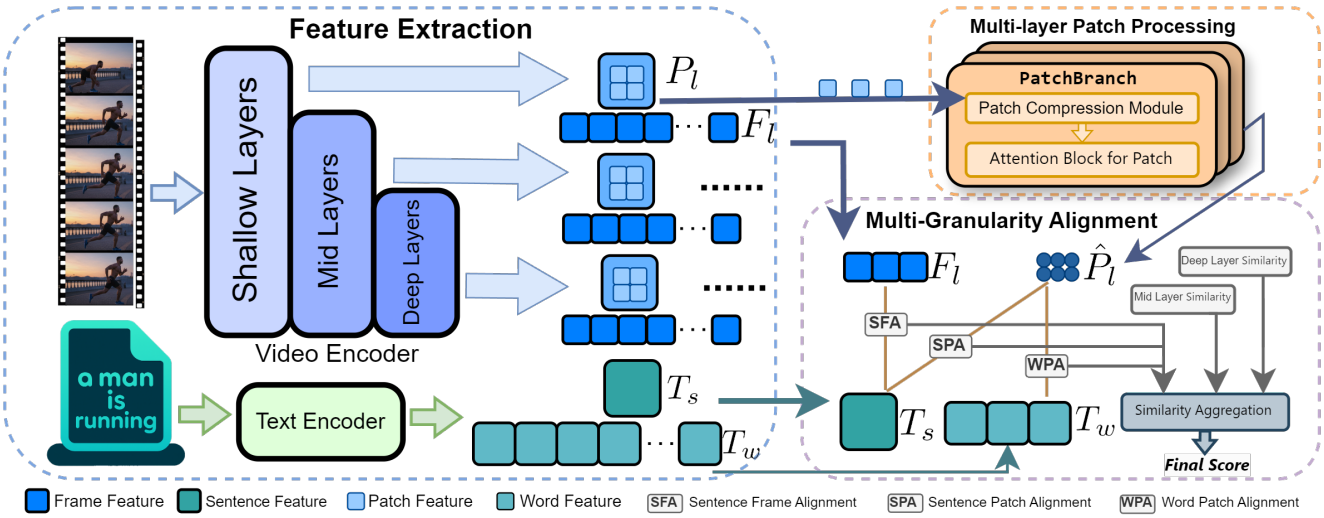

图 2。HVP-Net 的架构。我们首先从视频编码器中提取层次化帧（F_l）和图块（P_l）特征。然后，多层补丁处理模块将图块特征细化为 P_l。最后，我们在文本与多层视频特征（F_l、 P_l）之间进行多粒度对齐，并聚合逐层相似度分数以进行检索。

$T_s \in \mathbb{R}^D$ 和词级特征 $T_w \in \mathbb{R}^{L_w \times D}$。

不同于依赖最后一层输出的先前工作，我们从视觉变换器的多个中间层 $\mathcal{L}$ 中提取特征。对于包含 N 帧的视频，每一层 $l \in \mathcal{L}$ 都会提供层次化特征：帧特征 $F_l \in \mathbb{R}^{N \times D}$ （来自 [CLS] 标记）和图块特征 $P_l \in \mathbb{R}^{N \times M \times D}$ （来自图块标记，每帧包含 M 个图块）。随后，这种由粗到细的表示将输入我们的细化模块。

2.2. 多层图块处理（MPP）

为从冗余图块特征中提炼语义概念，我们引入**多层补丁处理模块（MPP）**，该模块迭代细化来自每一层的图块特征 P_l。每次迭代都包含一个“压缩—细化”循环：

1.**图块压缩模块** 我们首先计算所有 M 图块标记的显著性分数。随后，密度峰值聚类（DPC）算法从这些标记中动态识别 K 聚类中心（$K \ll M$）。最后，根据聚类分配，通过显著性加权平均将标记合并为 K 特征。

2. **图块注意力块：** K 压缩特征在交叉注意力块中充当查询，关注原始 M 标记，以细粒度细节丰富压缩表示。

ˆ这一迭代过程为每一层输出一个经过精炼的图块特征集 $P_l \in \mathbb{R}^{N \times K \times D}$，形成视觉实体的蒸馏式多层级表示。

2.3. 多粒度对齐与优化

为确保检索精准，我们在三个语义粒度上进行跨模态对齐，涵盖所有 L 提取的特征层级。对于给定的文本-视频对 $(i,\ j)$，我们计算每一层 l 的以下相似度分数：

1. **句子-帧（SF）对齐：** 全局句子特征 T_{s_i} 与帧特征 F_{l_j}之间的相似度。2. **句子-图块（SP）对齐：** T_{s_i} 与精炼图块概念 P_{l_j}之间的相似度。3. **词-图块（WP）对齐：** 词特征T_{w_i} 与精炼图块概念P_{l_j}之间的相似度。

对于图块级和词级特征，最终相似度分数并非通过简单池化得到。相反，我们首先利用标记到图块的相似度矩阵上的最大池化，确定每个词标记对应的最相关图块（反之亦然）。随后，我们使用由小型多层感知机预测的可学习权重，对这些最大相似度分数进行加权聚合。这样，模型便能动态强调最显著的词和图块贡献。对于给定层 l，词到图块的相似度分量计算如下：

$$S^{(l)}_{\mathrm{w}\to\mathrm{p}}(i,j) = \sum_{k=1}^{N_w} W_{\mathrm{MLP}}(w_{ik}) \cdot \max_{m=1}^{N_p} \mathrm{sim}(w_{ik}, p_{jm}) \quad (1)$$

此外，还会计算对称的图块与词语相似度，最终得分为两者的平均值。

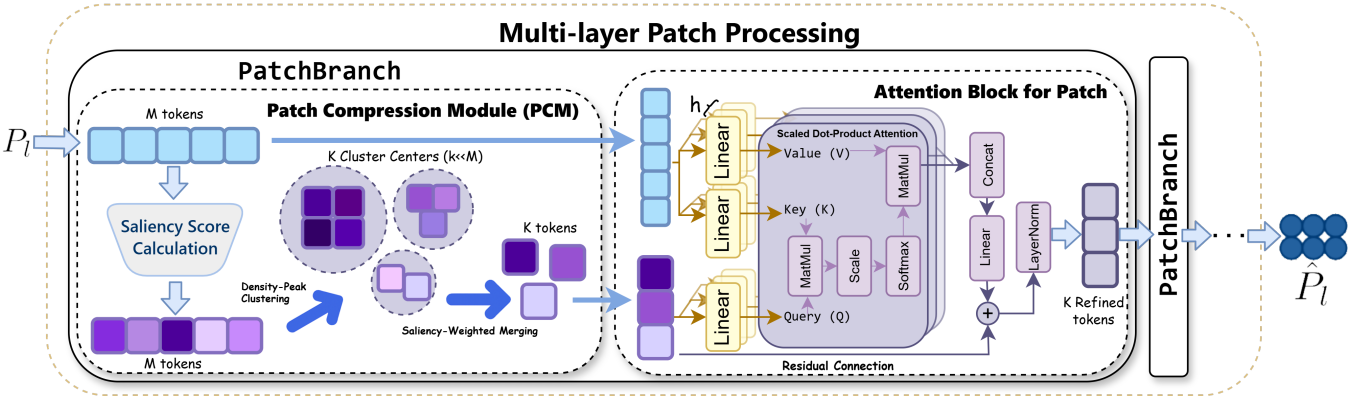

图 3。多层图块处理（MPP）模块的架构，该模块迭代优化输入图块特征 P_l，得到 $\hat{P}_l$。在每个处理块中，图块压缩模块（PCM）首先使用显著性引导的聚类，将 M 个词元提炼为 K 个核心概念。随后，通过交叉注意力机制对这些概念进行优化，该机制会回溯关注原始词元。

目标函数。我们使用对称对比损失（InfoNCE）[19]优化模型。总损失是每个特征层级上各对齐粒度对应的双向对比损失之和：

$$\mathcal{L}_{\text{total}} = \sum_{l=1}^{L} (\mathcal{L}_{\text{SF}}^{(l)} + \mathcal{L}_{\text{SP}}^{(l)} + \mathcal{L}_{\text{WP}}^{(l)}) \tag{2}$$

其中，每个组成部分（例如 $\mathcal{L}^{(l)}{}_{\text{SF}}$）都是基于第$l$层的句子-帧相似度矩阵计算的标准 InfoNCE 损失。这种跨多个层级和粒度的联合优化，确保学得一致且稳健的跨模态语义空间。

推理。在推理时，查询文本与候选视频之间的最终相似度分数，是所有粒度和层级上计算出的相似度分数之和，从而与训练目标保持一致。

3. 实验

我们进行了大量实验，以评估所提出的 HVP-Net。我们首先介绍实验设置，然后将我们的方法与最先进（SOTA）的方法进行比较，最后通过消融研究和定性示例进行深入分析。

3.1. 实验设置

数据集与评估指标。我们在三个主流基准上进行评估：**MSR-VTT** [20]，**DiDeMo** [21]，**ActivityNetCaptions** [22]。对于 MSR-VTT，我们采用标准的“1k-A”测试划分。我们使用标准检索指标：**K值召回率 (R@K, K=1, 5, 10)**（越高越好）、**中位排名 (MdR)**和**平均排名 (MnR)**（越低越好），分别评估文本到视频 (T2V) 和视频到文本 (V2T) 任务。

实现细节。我们的模型基于 CLIP ViT-B/32 骨干网络构建。我们以 224x224 的分辨率为每个视频均匀采样 12 帧，文本长度上限为 24 个词元。初始学习率设为 1e-4（CLIP 骨干网络的系数为 0.001），权重衰减设为 0.2。模型在 MSR-VTT 上训练 5 个 epoch，批次大小为 32。MPP 模块处理从视觉编码器的三个层提取的特征：浅层（第 1 层）、中间层（第 6 层）和深层（第 12 层）Transformer 层。

3.2. 与最先进方法的比较

我们将 HVP-Net 与近期的最先进方法进行比较，包括强大的、基于 CLIP 的模型，如 CLIP4Clip [15], X-CLIP[17],BiHSSP[23], MUSE[24]。

表1。在 MSR-VTT 1k-A 测试集上的性能比较。R@K 的单位为百分比（%）。最佳结果以**粗体**显示。

方法	R@1 ↑	R@5 ↑	R@10 ↑	中位数排名 ↓	平均排名 ↓
文本到视频（*T2V*）					
CLIP4Clip [15]	44.5	71.4	81.6	2.0	15.3
X-CLIP [17]	49.3	75.8	84.8	2.0	12.2
BiHSSP[23]	48.1	74.0	84.1	2.0	12.1
MUSE[24]	50.9	76.7	85.6	1.0	10.9
HVP-Net（本文方法）	**56.7**	**83.9**	**90.6**	**1.0**	**5.3**
视频到文本（*V2T*）					
CLIP4Clip [15]	42.7	70.9	80.6	2.0	11.6
X-CLIP [17]	48.9	76.8	84.5	2.0	8.1
BiHSSP[23]	48.0	74.1	83.5	2.0	9.0
MUSE[24]	49.7	77.8	86.5	2.0	7.4
HVP-Net（我们的方法）	**50.5**	**80.5**	**91.2**	**2.0**	**4.4**

如表1所示，HVP-Net 在 MSR-VTT 上显著超越了所有竞争方法。在文本到视频检索任务中，我们的方法取得了 56.7% 的 R@1，比最强基线 MUSE 高出 5.8 个百分点。这一领先表现也体现在

文本到视频和视频到文本任务的所有指标上。表2进一步证实了我们模型的优越性和泛化能力：它在 DiDeMo 和 ActivityNet 上均取得了新的最先进结果，在所有报告的召回率指标上都优于以往方法。

表2. 文本到视频检索在 DiDeMo 和 ActivityNet 上与最先进方法的性能比较。

方法	DiDeMo			ActivityNet		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Clip4Clip [15]	-	-	-	40.5	72.4	83.6
CLIP-VIP [25]	48.6	77.1	84.4	-	-	-
HBI [26]	-	-	-	42.2	73.0	84.6
DiffusionRet [27]	46.7	74.7	82.7	-	-	-
HVP-Net（我们的方法）	**57.1**	**83.1**	**87.0**	**43.5**	**78.3**	**88.2**

3.3. 消融实验与分析

我们在 MSR-VTT 数据集上进行了全面的消融实验，以分析 HVP-Net 的关键组成部分。表3报告了文本到视频任务的所有结果。

核心组件的有效性。 我们首先验证两项主要贡献：多层策略和多层图块处理模块。如表3（左）所示，仅使用最终层特征的基线模型取得了54.2%的R@1。完整的HVP-Net模型将性能提升至56.7%。值得注意的是，若不使用我们提出的MPP 模块而直接聚合原始中间特征（+多层（不含MPP）），性能会急剧下降至22.1%，这证实了净化过程对于有效融合多层信息至关重要。

层选择的影响。 我们在表3（中）中分析了不同的层选择策略。尽管使用连续的最后3层（55.6%）相比最后1层基线（54.2%）有所提升，但我们默认从不同语义阶段（1、6、12）进行采样的策略表现最佳（56.7%）。这支持了我们的假设：融合低层细节与高层语义是有益的。使用更多层（1、3、6、9、12）会导致性能下降（51.9%），表明我们的三阶段采样是一种高效且有效的选择。

对齐损失的影响。 最后，对每个损失组件分别进行消融（表3，右侧）进一步证实了它们的必要性。移除句子-帧（$\mathcal{L}_{SF}$）、句子-图像块（$\mathcal{L}_{SP}$）或词-图像块（$\mathcal{L}_{WP}$）对齐损失后，性能分别下降**8.5**、**3.5**和显著的**23.4** 个百分点。这表明，三个监督层级都至关重要，其中细粒度的词-图像块对齐是对稳健跨模态理解影响最大的组件。

4. 结论

本文针对文本-视频检索中的信息冗余这一关键挑战进行了研究，这是仅依赖最终层特征的传统方法所存在的局限，

表3。对核心组件、层选择策略和损失函数的消融实验（T2V R@1，%）。 Δ表示相对于完整模型的性能下降。

组件消融	R@1	层选择	R@1	损失消融	R@1（Δ）
基线（最后1层）	54.2	最后1层	54.2	完整模型	56.7 (-)
+ 多层（不含 MPP）	22.1	最后3层	55.6	不含 $\mathcal{L}_{SF}$	48.2 (-8.5)
+MPP（完整模型）	**56.7**	5层 1,3,6,9,12	51.9	不含 $\mathcal{L}_{SP}$	53.2 (-3.5)
		默认 1,6,12	**56.7**	不含 $\mathcal{L}_{WP}$	33.3 (-23.4)

我们提出了**HVP-Net**，这是一种新颖的框架，包含**多层补丁处理（MPP）模块**。该模块通过优化视觉编码器中间层的特征，有效提炼出显著概念，从而构建出更稳健的跨模态嵌入空间。我们的方法在多个基准测试中取得了新的最先进性能，其中在 MSR-VTT 上的 R@1 达到 56.7%，表现尤为突出。我们的工作验证了挖掘预训练模型内部层次特征是实现跨模态对齐的一种有效策略。未来的工作可以探索自适应层选择策略，以及提升视频到文本检索性能的具体机制。

5. 参考文献

[1]Zhichao Han、Azreen Bin Azman、Mas Rina Binti Mustaffa 和 Fatimah Binti Khalid，“跨模态检索：方法、数据集与未来展望综述”，*IEEE Access*，第 12 卷，第 115716–115741 页，2024 年。

[2] 方峰、王晓飞和李如实，“利用对应关系自编码器进行跨模态检索”，载于第*22*届 *ACM* 国际多媒体会议论文集，2014年，第7–16页。

[3] 谢泽群、季浩明、李承轩和孟令伟，“利用噪声对应关系进行文本人员搜索的动态不确定性学习”，2025年。

[4] 谢泽群、王楚欣、王业强、蔡思航、王树磊和金涛，“用于人员检索的聊天驱动文本生成与交互”，载于*2025*年自然语言处理实证方法会议论文集，2025年，第5259–5270页。

[5]柯宗、曹玉清、陈振瑞、尹宇辰、何守超和程宇，“通过多源数据预警加密货币反转风险”，《金融研究快报》，第107890页，2025年。

[6] ZongKe、JiaqingShen、Xuanyi Zhao、Xinghao Fu、Yang Wang、Zichao Li、Lingjie Liu 和 Huailing Mu，“一种具有 GRU-CNN-GA 融合的稳定技术特征”，《应用软计算》，第 114302 页，2025年。

[7]Weihai Lu、Yu Tong 和 Zhiqiu Ye，“Dammfnd：领域感知的多模态多视图假新闻检测”，载于《*AAAI* 人工智能会议论文集》，2025年，第 39 卷，第 559–567 页。

[8] Yu Tong、Weihai Lu、Xiaoxi Cui、Yifan Mao 和 Zhe- jun Zhao，《Dapt：用于多模态虚假新闻检测的领域感知提示调优》，载于第*33*届 *ACM*国际多媒体会议论文集，2025年，页码 7902–7911。

[9] Weihai Lu 和 Li Yin，《Dmmd4sr：基于扩散模型的序列推荐多层级多模态去噪》，载于第*33*届 *ACM*国际多媒体会议论文集，2025年，页码 6363–6372。

[10] Xiaoxi Cui、Weihai Lu、Yu Tong、Yiheng Li 和 Zhejun Zhao，《基于条件扩散的特征去噪实现多模态多行为序列推荐》，载于第*48*届 *ACM SIGIR*信息检索研究与开发国际会议论文集，2025年，页码 1593–1602。

[11] 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，《从自然语言监督中学习可迁移的视觉模型》，2021年。

[12] Xiaohan Wang、LinchaoZhu 和 YiYang，《T2vlad：用于文本-视频检索的全局-局部序列对齐》，2021年。

[13] Shaobo Min、Weijie Kong、Rong-Cheng Tu、Dihong Gong、Chengfei Cai、WenzheZhao、Chenyang Liu、Six-iao Zheng、Hongfa Wang、Zhifeng Li 和 Wei Liu，《混元文本视频检索：文本-视频检索》，2022年。

[14] Lisai Zhang、Hong Wu、Qing Chen、Yuan Deng、Jochen Siebert、Zili Li、Yuhan Han、Dong Kong 和 Zhe Cao，《Vldeformer：用于快速跨模态检索的视觉-语言分解式 Transformer》，知识型系统，第252卷，第109316页，2022年。

[15] Huaishao Luo、Lei Ji、Ming Zhong、Yang Chen、Wen Lei、NanDuan 和 Tianrui Li，“Clip4clip：面向端到端视频片段检索的 CLIP 实证研究”，2021年。

[16] XunYang、JianfengDong、Yixin Cao、Xun Wang、Meng Wang 和 Tat-Seng Chua，“面向复杂查询视频检索的树增强跨模态编码”，2020年。

[17]Yiwei Ma、Guohai Xu、Xiaoshuai Sun、Ming Yan、Ji Zhang 和 Rongrong Ji，《X-clip：用于视频文本检索的端到端多粒度对比学习》，2022年。

[18] Yao Lu、Wen Yang、Yunzhe Zhang、Zuohui Chen、Jinyin Chen、QiXuan、ZhenWang 和 Xiaoniu Yang，《利用图模块度理解深度神经网络的动态信息》，载于欧洲计算机视觉会议。施普林格，2022年，第225–242页。

[19] A¨aron vanden Oord、YazheLi 和 Oriol Vinyals，“基于对比预测编码的表征学习”，*arXiv*，卷号 abs/1807.03748，2018年。

[20] 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”，发表于*2016*年 *IEEE* 计算机视觉与模式识别会议（*CVPR*），2016年，第 5288–5296 页。

[21] Lisa Anne Hendricks、Oliver Wang、Eli Shechtman、Josef Sivic、Trevor Darrell 和 Bryan Russell，“利用自然语言定位视频片段”，2017年。

[22] Ranjay Krishna、Kenji Hata、Frederic Ren、李飞飞和 Juan Carlos Niebles，“视频中的密集字幕生成事件”，载于*2017*年*IEEE*国际计算机视觉大会 *(ICCV)*，2017年，第706–715页。

[23] Yang Liu、Shudong Huang、Deng Xiong 和 Jiancheng Lv，“通过双向层次滑动语义探针学习动态相似性，以实现高效的文本视频检索”，*AAAI*人工智能大会论文集，第39卷，第6期，第5667–5675页，2025年4月。

[24] Haoran Tang、Meng Cao、Jinfa Huang、Ruyang Liu、Peng Jin、Ge Li 和 Xiaodan Liang，“MUSE：Mamba 是一种用于文本-视频检索的高效多尺度学习器”，2025年。

[25] Hong Xue、Yuchong Sun、Bei Liu、Jianlong Fu、Rui Song、Houqiang Li 和Jiebo Luo，“CLIP-VIP：使预训练图像文本模型适配视频语言对齐”，发表于第十一届国际学习表征会议，2022年。

[26] Peng Jin、Jin Huang、Peng Xiong、Shizhe Tian、Chang Liu、Xiang Ji、Li Yuan 和 Jian Chen，“视频文本作为博弈参与者：用于跨模态表征学习的层次化班兹哈夫交互”，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第2472–2482页。

[27] Peng Jin、Hao Li、Zekai Cheng、Ke Li、Xiang Ji、Chang Liu、Li Yuan 和 Jian Chen，“DiffusionRet：基于扩散模型的生成式文本-视频检索”，发表于*IEEE/CVF*国际计算机视觉会议论文集，2023年，第2470–2481页。