/TemplateVersion (2025.1)

含糊限制的文本-视频表示学习用于部分相关视频检索

Cheol-Ho Cho, WonJun Moon, Woojin Jun, MinSeok Jung, and Jae-Pil Heo*

Sungkyunkwan University

{ hoonchcho, wjun0830, junwoojinjin, minseokjung0328, jaepilheo } @gmail.com

Abstract

部分相关视频检索（PRVR）旨在检索视频中与给定文本查询相关的特定片段。PRVR 的典型训练过程假定每个文本查询仅与一个视频相关的一对一关系。然而，我们指出了文本和视频内容之间在其概念范围内的固有模糊性，并提出了一个将这种模糊性纳入模型学习过程的框架。具体来说，我们提出了模糊性抑制表示学习（ARL）来处理模糊的文本-视频对。首先，ARL 基于两个标准检测模糊对：不确定性和相似性。不确定性表示实例是否包含数据集中常见的共享上下文，而相似性表示成对的语义重叠。然后，针对检测到的模糊对，我们的 ARL 通过多正对比学习和双三元组边际损失分层学习语义关系。此外，我们深入研究视频实例内的细粒度关系。与典型的文本-视频级别训练不同的是，我们在同一个未剪辑视频的帧中处理固有的模糊性，其中通常包含多个上下文。这使我们能够在文本-帧级别进一步提升学习。最后，我们提出跨模型模糊性检测，以减轻在单个模型用于检测其训练中模糊对时发生的错误传播。结合所有组件，我们提出的方法在 PRVR 中展示了其有效性。

介绍

随着社会的进步，使用视频媒体进行信息传播已经变得普遍。因此，允许用户使用文本查询找到所需视频的文字到视频检索（T2VR）领域也受到了关注。然而，现有的 T2VR 方法常常假设视频只包含与文本查询相关的部分。这个假设与现实世界中的场景不符，因为视频的长度和上下文可以有所不同。为了解决这个问题，提出了部分相关的视频检索（PRVR）(Dong et al. 2022)，以处理那些只有特定视频片段与文本查询对应的未剪辑视频。

MS-SL (Dong et al. 2022) 最初建模了多尺度的视频特征，以便为文本查询可能涵盖的各种上下文做准备。另一方面，GMMFormer (Wang et al. 2024) 批评了 MS-SL 对多尺度片段的详尽建模的低效性，并建议使用高斯注意力仅对局部上下文进行编码。尽管这些方法在 PRVR 方面取得了显著进展，但文本视频对的标签含糊不清问题尚未得到探讨。

通常，由于探索所有文本和视频实例之间关系的成本很高，文本和视频实例通常以成对的方式标记。这导致以前的工作仅将配对的实例学习为正向关系，而将所有其他情况视为负向关系，即便在存在相似的视频文本对

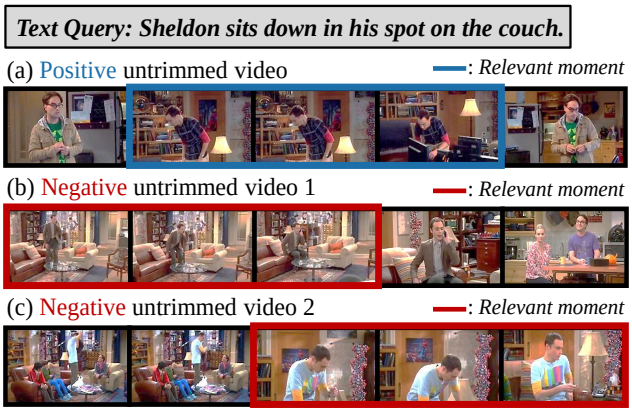


Figure 1: 文字与视频对之间的模糊关系示例。对于文本查询“谢尔顿坐在沙发上的他的位置。”，所有三个未剪辑的视频都包含相关场景。然而，只有视频 (a) 被学为正例，而视频 (b) 和 (c) 在先前的技术中通常被视为负例。这样的文字与视频之间的模糊关系更可能发生在未剪辑视频中，因为这些视频通常包含多样的背景。

的情况下也是如此。然而，我们认为成对标记的检索数据集往往会在文本视频实例之间引入模糊性。例如，如图 1 所示，虽然文本查询显然部分与顶部的配对视频相关，但它也与数据集中的其他视频实例相关。

在这方面，我们提出了一个名为模糊限制表示学习（Ambiguity-Restrained representation Learning, ARL）的框架，它利用模型的在线（每个周期）知识检测模糊关系中的实例。为了减少可能因将所有未配对的对视为负关系而产生的错误监督，这些模糊关系在目标中被考虑。为了确定模糊关系，我们使用不确定性和相似性测量，如图 2 所示。简单来说，我们将文本-视频对定义为处于模糊关系中，如果它们表现出高度的不确定性和高对间相似性，这表明由于它们在整个数据集中拥有共同的语义，并且彼此相似，这种对不能简单地被定义为负关系。我们提出的 ARL 在训练中包括了这些识别出的模糊关系。特别是，我们给予模型灵活性，通过放宽限制来处理具有模糊关系的实例，而正负关系则以传统方式学习。除了文本-视频关系，我们还进一步探索了具有模糊感知目标的文本-帧关系，因为未剪辑的视频通常包含多个上下文。最后，我们采用跨模型的模糊检测来减轻在检测模糊关系时可能发生的误差传播风险，因

*Corresponding author

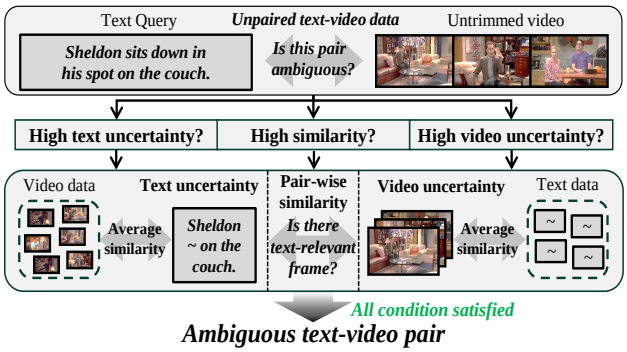


Figure 2: 模棱两可对检测的示例。为了识别模棱两可的文本-视频对，我们使用两个关键指标：不确定性和相似性。每个文本和视频实例的不确定性通过测量不同模态之间的平均相似性来计算。这反映了数据集中上下文重叠的程度。相似性则由文本与未剪辑视频中的帧之间的最大相似性来决定。当不确定性和相似性都很高时，我们定义该对具有模棱两可的关系，这体现了数据集中以及对之间共享的上下文。

为模型预测并使用其预测进行训练。

我们的主要贡献是：(1) 我们提出了消歧表征学习（ARL），这是第一个解决 PRVR 中标签模糊性的方法。通过建模实例之间的关系，ARL 减轻了学习不确定关系的影响。(2) 我们将 ARL 扩展到文本-帧级别，以处理未剪辑视频中的多重语境，增强学习过程中对所有帧的有效利用。(3) 我们引入交叉模型模糊检测，避免反复学习错误检测的模糊。(4) 我们在两个数据集上实现了最先进的性能，即 TVR 和 ActivityNet。

相关工作

文本到视频检索

文本到视频检索（T2VR）旨在通过对齐视频内容和文本描述来实现无元数据的搜索范式。为了从视觉-语言对齐模型中获益，通常使用在大规模文本-图像对上预训练的 CLIP 模型作为初始学习点。随后，为了解决视频和文本中信息量的不匹配问题，一些工作集中于设计特征匹配的基本单元 (Gorti et al. 2022; Lin et al. 2022)，例如，帧-词 (Wang et al. 2022) 和帧-句子 (Jin et al. 2023; Wu et al. 2023)。

在 T2VR 的文献中已经讨论了上下文范围的不确定性。特别是，已经有一些方法 (Fang et al. 2023; Li et al. 2024; Song and Soleymani 2019) 来解决这种不确定性。PVSE (Song and Soleymani 2019) 和 UATVR (Fang et al. 2023) 提取了文本-视频对的多方面表示。PAU (Li et al. 2024) 通过确保不同不确定性测量之间的一致性，解决了文本文本-视频数据中固有的偶然性不确定性。我们的研究在解决文本-视频数据不确定性这个高层次概念上有共同点。然而，我们的工作不同之处在于，关键焦点是探索所有文本-视频对之间的模糊关系。

部分相关视频检索

在超越 T2VR 的情境之外，PRVR 进一步针对搜索引擎的细粒度能力进行优化，即使只有部分上下文与给定的文本查询相对应时也能检索视频。为了解决 PRVR，典型的方法是将视频剪辑成多个片段。MS-SL (Dong et al. 2022) 全面构建了不同长度的剪辑，并与文本查询进行相似度匹配。GMMFormer (Wang et al. 2024) 在 attention 层中应用正态分布的权重，实现了在形成剪辑表示时的局部特征。

另一方面，我们的工作重点是检测由于视频检索数据集中的一对一标注而导致的文本-视频实例之间的模糊关系。

由于其实用性，带有噪声标签的学习引起了广泛关注 (Han et al. 2018; Li, Socher, and Hoi 2020; Azadi et al. 2015; Wang et al. 2019)。不确定性估计和协同训练框架是另两个受欢迎的研究方向。不确定性常用于检测噪声标签 (Neverova, Novotny, and Vedaldi 2019; Ju et al. 2022; Northcutt, Jiang, and Chuang 2021; Zheng and Yang 2021)，而协同训练框架在优化噪声标签方面被证明是有效的 (Han et al. 2018; Wei et al. 2020; Tan et al. 2021; Li, Socher, and Hoi 2020)。最近，这个问题也在视频文本学习的背景下得到了解决 (Lin et al. 2024)。为了应对成对标记关系中可能存在的模糊性，我们的研究采用了不确定性和协同训练的概念。

方法

概述

模糊受限表示学习（ARL）的概述如图 3 所示。为说明起见，我们计算训练集中所有文本和视频之间的相似性，以在每个迭代中定义文本-视频的不确定性。随后，批量索引的不确定性与小批量中的文本和视频之间的相似性一起传递到标签模糊检测（LAD）模块中。LAD 在两个层次上识别文本和视频模态之间的模糊关系：即文本-视频和文本-帧。最后，我们采用在 PRVR 文献中常用的双分支结构进行跨模型模糊检测。对跨模型模糊检测的每个分支采用相同的结构和输入，跨模型模糊检测使每个模型能够学习来自另一个模型的检测到的模糊集合。请注意，跨模型模糊检测未在图 3 中描述。为了更清晰地说明，本文的其余部分假设所有文本查询和视频实例都由相同数量的元素组成，即 L_q 个词和 L_v 帧。

给定训练数据集中 N_q 个查询中的第 i 个文本查询，我们使用预训练的文本编码器来提取每个词的特征。随后，我们使用一个全连接 (FC) 层和 ReLU 激活将词特征嵌入到低维空间中。在此之后，我们将位置编码加入到这些特征中，并使用一个 transformer 层来获得 d 维的词特征向量 $Q_i \in \mathbb{R}^{L_q \times d}$。最后，我们对词特征向量应用注意力池化模块，以获得查询文本嵌入 $q_i \in \mathbb{R}^d$。

给定 j -th 未修剪视频中的 N_v 训练集视频，我们使用预训练的 2D 或 3D CNN 提取帧特征 $V'_j \in \mathbb{R}^{L_v \times d_v}$。和文本分支对称，视频特征也通过带 ReLU 激活功能的全连接层以减少维度。之后，我们将位置编码 P 合并到提取的特征中，然后通过一个 transformer 层转发特征以获得 $V_j \in \mathbb{R}^{L_v \times d}$：

$$V_j = [v_{j1}, v_{j2}, ..., v_{jL_v}] = \text{Transformer}(\text{FC}(V'_j) + P), \quad (1)$$

其中 v_{jk} 指的是在 j -th 视频中的 k -th 帧特征。

相似性度量 给定文本和视频表示形式，帧级相似度分数 s^f 在文本查询特征 q_i 和视频帧特征 v_{jk} 之间被表示为：

$$s^f(q_i, v_{jk}) = \cos(q_i, v_{jk}), \quad (2)$$

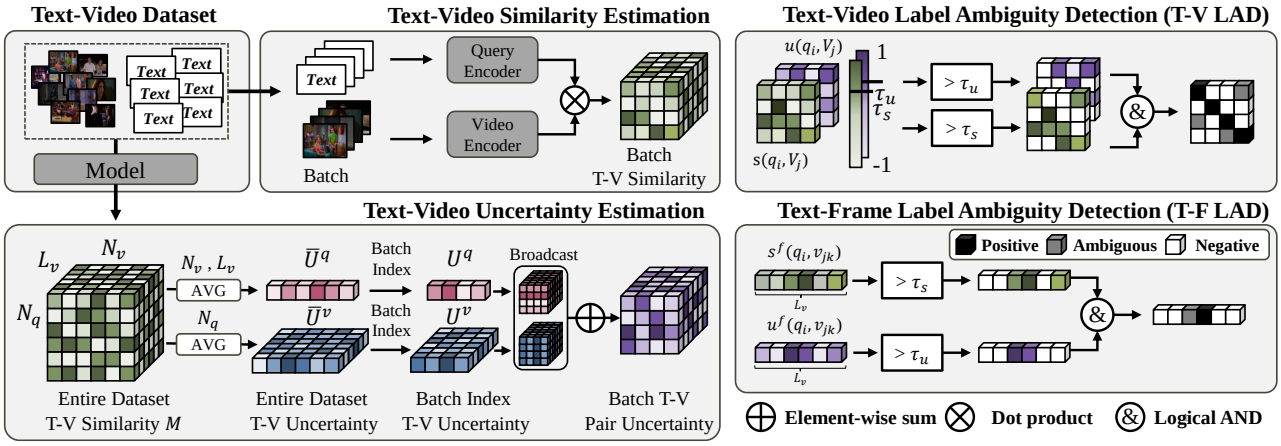


Figure 3: ARL 概述。(左) 给定文本-视频训练集，我们最初计算文本-视频相似性以在数据集层面计算不确定性。批次索引的不确定性结合批次中文本和视频之间的相似性用于探索小批次中的模糊文本-视频关系。(右) 两级标签模糊检测（LAD）模块检测模糊关系。文本-视频 LAD 采用 $s(q_i, V_j)$ 和 $u(q_i, V_j)$，即每个小批次中文本查询和视频之间的相似性和不确定性图。文本-帧 LAD 利用 $s^f(q_i, v_{jk})$ 和 $u^f(q_i, v_{jk})$，即每个文本-视频对中每个查询和视频帧之间的相似性和不确定性图。

其中 cos $(\cdot,\cdot)$ 是文本查询和视频帧之间的余弦相似度。然后，由于只有部分视频帧与文本相关，使用文本查询与视频帧之间的最大相似度值作为检索分数。检索（相似度）分数如下获得：

$$s(q_i, V_j) = s^f(q_i, v_{j\hat{k}}), \tag{3}$$

其中 $v_{j\hat{k}} = \text{argmax}_{v_{jk}} \cos(q_i, v_{jk})$ 和 $\hat{k}$ 分别表示具有最大相似度的视频帧及其索引。

不确定性限制的表示学习

歧义的定义。部分相关视频检索（PRVR）的典型数据集由成千上万的匹配文本视频对组成。一种流行的做法是将配对的文本视频数据视为正对，而将所有未配对的数据视为负对。然而，我们质疑将所有未配对数据视为负对是否正确。

我们通过为难以简单地分类为负例的未配对文本-视频对定义一种模糊关系来解决这个问题。具体来说，我们使用不确定性和相似性度量来识别这些模糊关系，并为模型在处理这些实例时提供灵活性。首先，不确定性表明每个实例是否包含常见的共享语义。它被计算为整个数据集的平均相似度，因此，如果许多实例与某个特定实例具有很高的相似度，则该实例的不确定性被认为较高。其次，相似性指的是文本-视频对的相似程度，定义为单个文本查询与视频帧表示之间的最大相似度。因此，高相似度的对表明某个视频实例包含一个帧，与给定的查询共享相似的属性。尽管这两个度量源自相同的相似操作，但我们注意到这两个度量作为不同标准有不同的目标。我们还注意到，我们在图 4 中的研究进一步验证了这两个度量呈现不同的分布。

我们的方法利用模型的在线知识（每个 epoch）来识别文本和视频之间的模糊关系。因此，我们需要将模型预热几个 epoch，以初步训练模型学习一般的文本-视频关系。在预热阶段，我们采用检索任务中常用的三元组排序损失 (Dong et al. 2023; Faghri et al. 2018) 和 infoNCE 损失 (Ma et al. 2022; Zhang et al. 2021)。

不确定性估计 在训练之前，通过考虑整个数据集的相似性来测量每个文本/视频实例的不确定性。具体来说，我们使用在线模型计算训练数据集中所有文本查询与视频帧之间的特征相似性图 $M \in \mathbb{R}^{N_q \times N_v \times L_v}$。这里，$M_{xyz}$ 表示整个数据集中第 x 个文本查询和第 z 帧第 y 个视频之间的相似性。通过相似性图 M，我们定义每个文本查询 $U^q \in \mathbb{R}^{N_q}$ 的数据集范围不确定性为该文本与数据集中所有视频帧之间的平均相似性，视频帧 $\bar{U}^v \in \mathbb{R}^{N_v \times L_v}$ 的不确定性为对所有文本查询的平均相似性：

$$\bar{U}^q_x = \frac{1}{N_v L_v} \sum_{y=1}^{N_v} \sum_{z=1}^{L_v} M_{xyz}, \quad \bar{U}^v_{yz} = \frac{1}{N_q} \sum_{x=1}^{N_q} M_{xyz}, \tag{4}$$

，其中 $\bar{U}^q_x$ 是第 x 个文本查询的不确定性值，$\bar{U}^v_{yz}$ 是第 y 个视频实例的第 z 帧的不确定性值。与视频帧平均相似性更高的文本和与查询平均相似性更高的帧都表现出更大的不确定性。

请注意，不确定性较高的实例意味着它的上下文可能与其他实例常常共享。这些相似性图和不确定性在每个时期都用等式 4 更新。

为了更清晰地索引每个小批次中的不确定性 U^q 和 U^v，我们将批次级别的不确定性子集定义为 U^q 和 U^v。请注意，U^q 和 U^v 是不确定性子集 $\bar{U}^q$ 和 $\bar{U}^v$ 的一部分，其中包括小批次中所有查询和视频的不确定性。

在下面，我们将文本查询和每个视频帧之间的不确定性值定义为 u^f，将查询和整个视频之间的不确定性定义为 u：

$$u(q_i, V_j) = \frac{1}{2}(U^q_i + U^v_{j\hat{k}}); u^f(q_i, v_{jk}) = \frac{1}{2}(U^q_i + U^v_{jk}), \tag{5}$$

，其中 $\hat{k}$ 表示与 i 文本查询最相似的帧索引，如公式 3 所述。

文本-视频标签歧义检测 为了在每个文本或视频实例中发现处于不明确关系的对，我们利用计算的不确定性 u 和相似度得分 s 。对于 i 次文本查询 q_i ，一组不明确的视频对被收集为：

$$\mathcal{A}_i^q = \{V_a \mid s(q_i, V_a) > \tau_s \text{ and } u(q_i, V_a) > \tau_u\}, \quad (6)$$

其中 τ_s 和 τ_u 是阈值超参数。

另一方面，视频 V_j 的一个模糊查询集定义如下：

$$\mathcal{A}_j^v = \{q_a \mid s(q_a, V_j) > \tau_s \text{ and } u(q_a, V_j) > \tau_u\}. \quad (7)$$

模糊感知表示学习。在训练过程中，我们利用边距三元组排序损失和对比学习 (Chen et al. 2020) ，沿袭之前的研究 (Dong et al. 2022; Wang et al. 2024) 。下面，我们列举了在目标上的修改，以实现在模糊关系中的应用。在对比学习的情况下，我们对监督对比学习 (Khosla et al. 2020) 进行了修改，以适应多正对比目标：

$$\mathcal{L}_{ij}^{\text{t2v}} = -\log\left(\frac{e^{s(q_i,V_j)} + \sum_{V_a \in \mathcal{A}_i^q} e^{s(q_i,V_a)}}{e^{s(q_i,V_j)} + \sum_{V \in \mathcal{A}_i^q \vee \mathcal{N}_i^q} e^{s(q_i,V)}}\right) \quad (8)$$

$$\mathcal{L}_{ij}^{\text{v2t}} = -\log\left(\frac{e^{s(q_i,V_j)} + \sum_{q_a \in \mathcal{A}_j^v} e^{s(q_a,V_j)}}{e^{s(q_i,V_j)} + \sum_{q \in \mathcal{A}_j^v \vee \mathcal{N}_j^v} e^{s(q,V_j)}}\right) \quad (9)$$

$$\mathcal{L}^{\text{nce}} = \frac{1}{n} \sum_{(q_i,V_j)\in\mathcal{B}} \mathcal{L}_{ij}^{\text{t2v}} + \mathcal{L}_{ij}^{\text{v2t}}, \quad (10)$$

其中 $\mathcal{B}$ 表示一个小批量，而 (q_i , V_j) 表示这一批次中的正对。$\mathcal{N}_j^v$ 和 $\mathcal{N}_i^q$ 是每个视频和查询的负样本集合，其中 $\mathcal{N}_j^v$ 包含既不与 j -th 视频有正向关系也不在模糊关系中的样本。简而言之，我们的多正对比目标通过容纳模糊关系让模型学习更加灵活。虽然在分子中的模糊集合内的实例不训练为负样本，但并非所有实例都需要被训练与锚点有正向关系。这是因为最大化单一相似性值仍然可以促进损失的收敛。

另一方面，我们为边缘三元组排序损失组织双三元组；一个包含 $\mathcal{L}_a^{\text{trip}}$ 的模糊集合，另一个包含 $\mathcal{L}_n^{\text{trip}}$ 的负样对，如下所示：

$$\mathcal{L}_a^{\text{trip}} = \frac{1}{n} \sum_{(q_i,V_j)\in\mathcal{B}} \{\max(0, m_a + s(q_a,V_j) - s(q_i,V_j)) + \max(0, m_a + s(q_i,V_a) - s(q_i,V_j))\} \quad (11)$$

$$\mathcal{L}_n^{\text{trip}} = \frac{1}{n} \sum_{(q_i,V_j)\in\mathcal{B}} \{\max(0, m + s(q_n,V_j) - s(q_i,V_j)) + \max(0, m + s(q_i,V_n) - s(q_i,V_j))\}, \quad (12)$$

其中 m_a 和 m 表示各自的边缘，$q_a \in \mathcal{A}_j^v$ 和 $V_a \in \mathcal{A}_i^q$ 分别代表每个视频和查询的模糊样本。q_n and V_n 是每个查询和视频的负样本。我们将 $\mathcal{L}_a^{\text{trip}}$ 的边缘设得更小，确保模糊实例比上下文无关实例 ($m_a < m$) 更加类似于锚点。通过减轻对模糊实例的距离约束，我们允许具有潜在正关系的模糊集合不参与负训练。边缘 m_a 用于维护层级，确保配对实例在方程 8 - 9 中作为正样本进行训练。综上所述，文本-视频对的模糊约束目标被表述为：$\mathcal{L}^{\text{video}} = \lambda_{\text{nce}}\mathcal{L}^{\text{nce}} + \mathcal{L}_a^{\text{trip}} + \mathcal{L}_n^{\text{trip}}$ ，其中 λ_{nce} 是一个用于平衡损失的超参数。

未剪辑视频中的文本帧标签歧义检测。对于未剪辑的视频，不同的上下文可能存在于同一个实例中。然而，探索相同视频内的关系仍然是一个未被探索的问题。因此，我们深入研究了文本查询与逐帧表示之间的关系。与在小批量中发现文本和视频实例之间的模糊关系的过程对称，我们在查询特征 q 与视频帧特征 v 之间应用了方程 6 和方程 7 。注意，相似度和不确定性衡量，即 s 和 u ，也分别被帧级相似度 s^f （方程 2 ）和不确定性 u^f （方程 5 ）所替代。因此，学习方程 8 至 12 中文本-视频关系的相同目标被用于学习每个视频的文本-帧关系。文本-帧对 $\mathcal{L}^{\text{frame}}$ 的目标也与 $\mathcal{L}^{\text{video}}$ 相同。我们注意到 λ_{nce} 是共享的。

跨模型歧义检测

可以使用模型自身的预测来检测模糊对，类似于自训练 (Balcan, Beygelzimer, and Langford 2006; Freund, Schapire, and Abe 1999) 。然而，我们指出，当模型依赖于其知识并逐步加强其初始不完善的预测时，模糊检测的错误传播易受攻击。

为解决这一挑战，我们利用两个相同的编码器，互相传递一个编码器检测到的模糊集合给另一个编码器，以减轻噪声标签的影响 (Han et al. 2018) 。给定两个模型，分别表示为 θ 和 Φ ，每个模型根据公式 4 -公式 7 使用其在线知识计算模糊的文本-视频对，并将其作为训练指导提供给另一个模型。

最后，预测的检索分数是基于每个模型的公式 3 的平均值得出的：

$$s(q_i, V_j) = \frac{1}{2}\big(s_\theta(q_i, V_j) + s_\Phi(q_i, V_j)\big), \quad (13)$$

其中 s_θ 和 s_Φ 分别表示来自模型 θ 和 Φ 的检索分数。

实验

数据集和指标

我们在两个大规模视频数据集上评估我们的方法，即 TVR (Lei et al. 2020) 和 ActivityNet Captions (Krishna et al. 2017) 。我们采用之前工作提供的数据划分 (Zhang et al. 2020, 2021) 。按照 (Dong et al. 2022) ，我们使用基于排名的召回率作为评估指标，即 R@K (K=1, 5, 10, 100)，其中 R@K 表示在排名列表前 K 项中成功检索到所需项目的查询比例。此外，我们报告所有召回率之和（SumR）以进行全面比较。

实现细节

我们使用 ResNet (He et al. 2016) 和 I3D (Carreira and Zisserman 2017) 用于 TVR，仅使用 I3D 用于 ActivityNet-Captions 来提取视觉特征。对于文本查询表示，我们对两个数据集都使用 RoBERTa (Liu et al. 2019) 特征。此外，为了展示我们的方法在大型模型上的有效性，我们进行了使用 CLIP-L/14 (Radford et al. 2021) 的实验。虽然典型的参数设置与 (Wang et al. 2024) 中相同，但阈值 τ_s 和 τ_u 在每个训练周期中使用训练数据集的相似性和不确定性分布值定义。特别地，τ_s 设置为正样本对的相似性分布的平均值，τ_u 设置为训练数据集的不确定性分布平均值所对应的值。更多细节在附录中提供。

Model	TVR					ActivityNet Captions				
	R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR
VCMR methods without moment localization										
XML (Lei et al. 2020)	10.0	26.5	37.3	81.3	155.1	5.3	19.4	30.6	73.1	128.4
ReLoCLNet (Zhang et al. 2021)	10.7	28.1	38.1	80.3	157.1	5.7	18.9	30.0	72.0	126.6
CONQUER (Hou, Ngo, and Chan 2021)	11.0	28.9	39.6	81.3	160.8	6.5	20.4	31.8	74.3	133.1
PRVR models										
MS-SL (Dong et al. 2022)	13.5	32.1	43.4	83.4	172.4	7.1	22.5	34.7	75.8	140.1
DL-DKD (Dong et al. 2023) †	14.4	34.9	45.8	84.9	179.9	8.0	25.0	37.4	77.1	147.6
GMMFormer (Wang et al. 2024)	13.9	33.3	44.5	84.9	176.6	8.3	24.9	36.7	76.1	146.0
Ours	15.6	36.3	47.7	86.3	185.9	8.3	24.6	37.4	78.0	148.3

Table 1: Resnet、I3D 和 Roberta 特征的性能比较。† 表示使用了额外的 CLIP-B/32 模型。

Model	TVR					ActivityNet Captions				
	R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR
MS-SL (Dong et al. 2022)	31.9	57.6	67.7	93.8	251.0	14.7	37.1	50.4	84.6	186.7
GMMFormer (Wang et al. 2024)	29.8	54.2	64.6	92.5	241.1	15.2	37.7	50.5	83.7	187.1
Ours	34.6	60.4	70.7	94.4	260.1	15.3	38.4	51.5	85.2	190.4

Table 2: 在 TVR 和 ActivityNet Captions 上的 CLIP-L/14 特征性能比较。

	MS-SL	GMMFormer	Ours
FLOPs (G)	1.29	1.95	1.23
Params (M)	4.85	12.85	5.34

Table 3: 模型复杂度比较。

inference runtime (ms)				
Video size	1000	1500	2000	2500
MS-SL	0.366	0.606	0.759	0.893
GMMFormer	0.264	0.267	0.270	0.293
Ours	0.294	0.391	0.427	0.612

Table 4: 推理时间比较。

与最先进技术的比较

检索性能。我们报告了将我们的方法与最先进的视频语料库时刻检索和 PRVR 方法进行比较的结果。请注意，所有性能都是在没有使用时刻监督的情况下获得的。在表 1 和表 2 中，我们展示了 TVR 和 ActivityNet Captions 数据集上的结果。如所观察到的，我们提出的方法在所有召回指标上都优于以前的工作，相较于使用 ResNet、I3D 和 Roberta 的 GMMFormer，在 SumR 中分别实现了 9.3 % 和 2.3 % 的差距。此外，我们强调了在 CLIP-B/32 模型中通过知识蒸馏（标记为 † ）进行增强的情况下，我们和 DL-DKD (Dong et al. 2023) 之间的一致差距。我们将这一卓越的性能归因于对 PRVR 中通过未裁剪视频进行一对一关系学习中的不明确性的研究。这一趋势在使用 CLIP-L/14 时也得到同样的观察，并分别相较于 MS-SL 和 GMMFormer 表现出 9.1 % 和 3.3 % 的提高。

复杂度分析。在本节中，我们展示了模型复杂度分析，如表 3 和 4 所示。我们分析了三个方面：FLOPs、参数数量，以及处理单个文本查询在 Nvidia RTX 3090 GPU 上所需的运行时间。虽然我们的方法在 FLOPs 和参数方面在 PRVR 方法中表现出较高的效率，但其运行时间相对比 GMMFormer 慢。FLOPs 和参数数量的效率归因于我们简化的 transformer 架构，而 GMMFormer 通过使用不同高斯核的多个注意力块并行实现。我们方法相对较慢的运行时间是由于缺少 GMMFormer 中使用的汇总帧级特征。然而，我们的方法仍然达到了实时运行时间少于 1 毫秒的性能，这代表了性能和运行速度之间一个良好平衡的权衡。

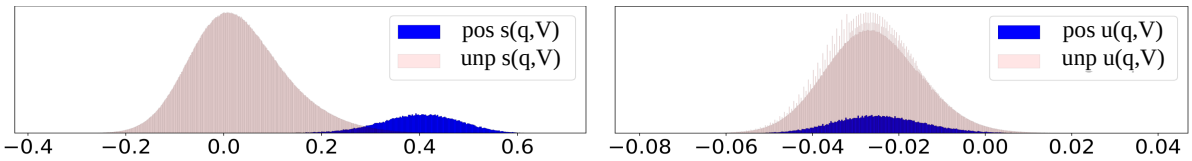
	T-V	T-F	C.L	R@1	R@5	R@10	R@100	SumR
(a)	-	-	-	32.8	58.1	68.2	93.7	252.8
(b)	✓	-	-	33.6	58.9	69.4	94.5	256.4
(c)	✓	✓	-	34.3	59.9	70.1	94.4	258.7
(d)	✓	-	✓	34.3	59.9	69.9	94.3	258.4
(e)	✓	✓	✓	34.6	60.4	70.7	94.4	260.1

Table 5: 消融研究用于调查不同组件对 TVR 的有效性。T-V 模糊性和 T-F 模糊性表示在文本-视频表示学习中使用模糊感知的表示学习和使用文本-帧表示学习。

消融及进一步研究

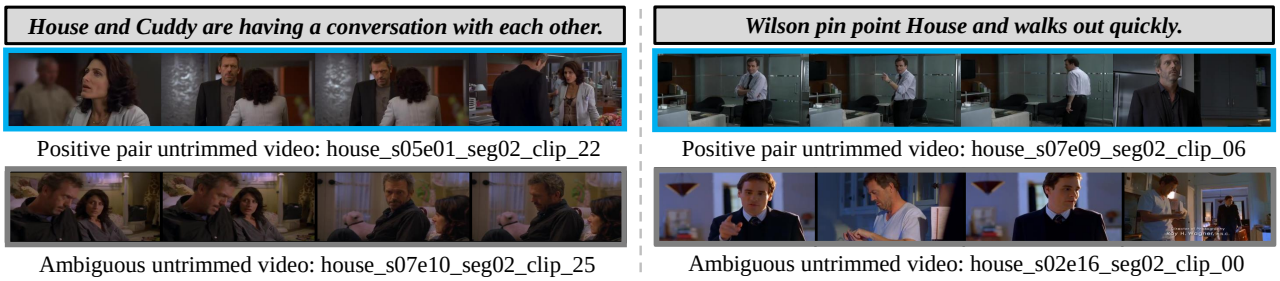
组件分析。为了理解每个组件的有效性，我们在表中进行了组件消融实验。通过比较行 (a) 和 (b)，我们的基础模型与在表示学习中通过训练处理文本-视频模糊的模型，我们观察到这在结果上是有效的，因为它使 SumR 增加了 3.6 个点。这表明在学习文本-视频关系时的灵活性对于 PRVR 是重要的，因为文本查询和视频在背景相似性方面通常表现出模糊性。此外，行 (c) 和 (e) 的增加突出显示了探索每个视频中的文本-帧关系的潜力。随后，跨模型学习被验证对 PRVR 是有益的，如行 (d) 和 (e) 所示。总之，我们组件的益处表明，在成对标记的数据集中仅考虑成对的文本-视频实例彼此在语境上相似容易形成文本-视频实例之间的模糊关系（尤其对于 PRVR）。

模糊集学习策略。在我们的工作中，我们赋予模型在学习模糊关系时的灵活性。然而，还有其他选项可以将模糊集中的所有文本-视频实例视为给定锚点的正例或将其从训练中排除。在表 6 中，我们报告了其他选项的性



(a) Similarity distribution on TVR (b) Uncertainty distribution on TVR

Figure 4: 视频问答任务中，正配对和未配对文本-视频对的相似性和不确定性分布。在图 (a) 中展示了相似性分布。正集合的分布通常高于负集合的分布。在图 (b) 中，展示了不确定性分布。由于不确定性值不太受单一对的相似性影响（对于正对），可见正集合和未配对的分布形成相似。



Positive pair untrimmed video: house_s05e01_seg02_clip_22

Positive pair untrimmed video: house_s07e09_seg02_clip_06

Ambiguous untrimmed video: house_s07e10_seg02_clip_25

Ambiguous untrimmed video: house_s02e16_seg02_clip_00

Figure 5: 在 TVR 数据集上的歧义检测结果。对于给定的查询，我们将经过训练过程后检测到处于歧义关系的未剪辑视频进行可视化（灰色方框）。这些视频虽然未与查询成对作为正实例，但与查询高度相关。

$\mathcal{A}$	R@1	R@5	R@10	R@100	SumR
Positive	34.0	59.7	70.1	94.5	258.3
Ignore	34.5	60.1	70.1	94.5	259.2
Ours	34.6	60.4	70.7	94.4	260.1

Table 6: 不同模糊集使用方式的性能比较。

能。具体来说，我们采用了一种监督对比目标（Khosla et al. 2020）来最大化每个模糊对之间的相似性以视为正例，或者利用掩码操作来忽略。由于在这两种情况下性能均下降，我们认为模糊集 $\mathcal{A}$ 是锚点的正负关系实例的混合。因此，这些结果表明简单地将模糊集定义为正例或负例可能导致次优结果。

不确定性 & 相似性分布。我们认为，用于检测歧义的这两个度量标准起着不同的作用，并表现出不同的分布。在图 4 中，我们展示了每个度量在 TVR 上的分布。（ActivityNet Captions 与 TVR 类似。）如图所示，相似性（左）和不确定性（右）的分布显示出不同的形状。具体来说，相似性上的正负分布是可区分的，表明正配对对之间的相似性通常高于未配对的文本-视频对。相反，我们观察到正样本和未配对集的不确定性值的分布相似，因为整个数据集的语义重叠程度不依赖于单个成对相似性。注意，绘图中使用的是文本-视频平均不确定性值，因为成对不确定性用于检测歧义对（方程 5）。这些结果表明，相似性和不确定性具有独立的意义，二者都应被同时考虑。

定性结果

我们进行了分析，以验证检测到的文本-视频对是否确实包含与一个锚点的模糊关系。在图 5 中，我们绘制了两个文本查询以及配对视频和模糊关系中的视频（用灰色框标注）。例如，我们观察到场景中的角色是相同的，同时整体环境也非常相似。特别是，左侧示例的影片包含可以用给定查询表达的时刻。这证实了我们的模糊性检测有效地捕捉到了模糊关系，并减少了将所有未配对的文本-视频对视为负集时发生的错误监督的影响。

结论

在本文中，我们解决了成对标注的文本视频数据中的文本与视频对之间关系模糊的问题。为了解决这一挑战，我们提出了”模糊受限表示学习”（Ambiguity-Restrained representation Learning，ARL），旨在减轻从文本与视频之间模糊关系中学习的影响。ARL 通过不确定性首先评估每个文本或视频是否可能在数据集中包括共同上下文。随后，计算每个小批次内的相似性以识别模糊的关系。这些关系随后被纳入模糊感知表示学习框架中，使模型在学习这些关系时具有灵活性。我们的结果表明，一对一关系学习易受到文本与视频之间模糊关系的影响。

本研究部分得到 MSIT/IITP（编号：2022-0-00680，2020-0-01821，2019-0-00421，RS-2024-00459618，RS-2024-00360227，RS-2024-00437102，RS-2024-00437633）和 MSIT/NRF（编号：RS-2024-00357729）的资助。

References

Azadi, S.; Feng, J.; Jegelka, S.; and Darrell, T. 2015. Auxiliary image regularization for deep cnns with noisy labels. arXiv preprint arXiv:1511.07069.

Balcan, M.-F.; Beygelzimer, A.; and Langford, J. 2006. Agnostic active learning. In Proceedings of the 23rd international conference on Machine learning, 65–72.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6299–6308.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In International conference on machine learning, 1597–1607. PMLR.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022. Partially Relevant Video Retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, 246–257.

Dong, J.; Zhang, M.; Zhang, Z.; Chen, X.; Liu, D.; Qu, X.; Wang, X.; and Liu, B. 2023. Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 11302–11312.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives.

Fang, B.; Liu, C.; Zhou, Y.; Yang, M.; Song, Y.; Li, F.; Wang, W.; Ji, X.; Ouyang, W.; et al. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. arXiv preprint arXiv:2301.06309.

Freund, Y.; Schapire, R.; and Abe, N. 1999. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780): 1612.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5006–5015.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

Hou, Z.; Ngo, C.-W.; and Chan, W. K. 2021. CONQUER: Contextual query-aware ranking for video corpus moment retrieval. In Proceedings of the 29th ACM International Conference on Multimedia, 3900–3908.

Jin, P.; Li, H.; Cheng, Z.; Huang, J.; Wang, Z.; Yuan, L.; Liu, C.; and Chen, J. 2023. Text-video retrieval with disentangled conceptualization and set-to-set alignment. arXiv preprint arXiv:2305.12218.

Ju, L.; Wang, X.; Wang, L.; Mahapatra, D.; Zhao, X.; Zhou, Q.; Liu, T.; and Ge, Z. 2022. Improving medical images classification with label noise using dual-uncertainty estimation. IEEE transactions on medical imaging, 41(6): 1533–1546.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. Advances in neural information processing systems, 33: 18661–18673.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, 706–715.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, 447–463. Springer.

Li, H.; Song, J.; Gao, L.; Zhu, X.; and Shen, H. 2024. Prototype-based Aleatoric Uncertainty Quantification for Cross-modal Retrieval. Advances in Neural Information Processing Systems, 36.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394.

Lin, C.; Wu, A.; Liang, J.; Zhang, J.; Ge, W.; Zheng, W.-S.; and Shen, C. 2022. Text-adaptive multiple visual prototype matching for video-text retrieval. Advances in neural information processing systems, 35: 38655–38666.

Lin, Y.; Zhang, J.; Huang, Z.; Liu, J.; Wen, Z.; and Peng, X. 2024. Multi-granularity correspondence learning from long-term noisy videos. arXiv preprint arXiv:2401.16702.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, 638–647.

Neverova, N.; Novotny, D.; and Vedaldi, A. 2019. Correlated uncertainty for learning dense correspondences from noisy labels. Advances in Neural Information Processing Systems, 32.

Northcutt, C.; Jiang, L.; and Chuang, I. 2021. Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 70: 1373–1411.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, 8748–8763. PMLR.

Song, Y.; and Soleymani, M. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1979–1988.

Tan, C.; Xia, J.; Wu, L.; and Li, S. Z. 2021. Co-learning: Learning from noisy labels with self-supervision. In Proceedings of the 29th ACM International Conference on Multimedia, 1405–1413.

Wang, Q.; Zhang, Y.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022. Disentangled Representation Learning for Text-Video Retrieval. arXiv:2203.07111.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF international conference on computer vision, 322–330.

Wang, Y.; Wang, J.; Chen, B.; Zeng, Z.; and Xia, S.-T. 2024. GMMFormer: Gaussian-Mixture-Model based Transformer for Efficient Partially Relevant Video Retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 5767–5775.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 13726–13735.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10704–10713.

Zhang, B.; Hu, H.; Lee, J.; Zhao, M.; Chammas, S.; Jain, V.; Ie, E.; and Sha, F. 2020. A hierarchical multimodal encoder for moment localization in video corpus. arXiv preprint arXiv:2011.09046.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Video corpus moment retrieval with contrastive learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 685–695.

Zheng, Z.; and Yang, Y. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. International Journal of Computer Vision, 129(4): 1106–1120.