Probabilistic Vision-Language Representation for Weakly Supervised Temporal Action Localization

Geuntaek Lim
Sejong University
Seoul, Republic of Korea
gtlim@rcv.sejong.ac.kr

Hyunwoo Kim
Sejong University
Seoul, Republic of Korea
hwkim@rcv.sejong.ac.kr

Joonsoo Kim
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
joonsookim@etri.re.kr

Yukyung Choi*
Sejong University
Seoul, Republic of Korea
ykchoi@rcv.sejong.ac.kr

ABSTRACT

Weakly supervised temporal action localization (WTAL) aims to detect action instances in untrimmed videos with only video-level annotations. As many existing works optimize WTAL models based on action classification labels, they encounter the task discrepancy problem (*i.e.*, localization-by-classification). To tackle this issue, recent studies have attempted to utilize action category names as auxiliary semantic knowledge with vision-language pre-training (VLP). However, there are still areas where existing research falls short. Previous approaches primarily focused on leveraging textual information from language models but overlooked the alignment of dynamic human action and VLP knowledge in joint space. Furthermore, the deterministic representation employed in previous studies struggles to capture fine-grained human motion. To address these problems, we propose a novel framework that aligns human action knowledge and VLP knowledge in the probabilistic embedding space. Moreover, we propose intra- and inter-distribution contrastive learning to enhance the probabilistic embedding space based on statistical similarities. Extensive experiments and ablation studies reveal that our method significantly outperforms all previous state-of-the-art methods. Code is available at https://github.com/sejong-rcv/PVLR

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Video Understanding, Human Action Understanding, Vision Language Pre-training

*Corresponding author.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681537

ACM Reference Format:
Geuntaek Lim, Hyunwoo Kim, Joonsoo Kim, and Yukyung Choi. 2024. Probabilistic Vision-Language Representation for Weakly Supervised Temporal Action Localization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681537

1 INTRODUCTION

The development of multimedia services, such as YouTube and Netflix, has sparked growing interest in the field of computer vision for analyzing long-form videos. Temporal Action Localization (TAL) refers to the problem of precisely determining the time intervals in a lengthy, untrimmed video when human activities occur, which is fundamentally significant for video understanding [15, 30, 31].

Fully supervised TAL [3, 4, 26, 27, 47, 56, 57] handles this task using rich frame-level annotations. Despite its success, training a TAL model with dense frame-level annotation poses challenges due to the high annotation costs and limited generality. To address these challenges, weakly supervised temporal action localization (WTAL) [9, 11, 18, 21, 24, 33, 44], which only requires video-level categorical labels, has received a lot of attention. In scenarios with only video-level category supervision, existing WTAL methods address the localization problem by selecting discriminative snippets[1] primarily for video-level classification. However, these classification-based approaches suffer from the task discrepancy problem inherent in a localization-by-classification framework. To tackle WTAL through a localization-by-localization framework, significant research [12, 25, 42, 61] has focused on snippet-level pseudo-label methods. Yet, these pseudo labels, constrained by video-level annotations, inherently carry noisy proposals and fail to achieve the desired accuracy.

Recently, to tackle these problems, some approaches [17, 24] have leveraged action category text information to guide powerful semantic knowledge without incurring additional annotation costs. These approaches establish additional learning cues by exploiting category text embedding vectors instead of merely utilizing category information as one-hot vectors. While significant improvements were made through additional semantic information, some factors were overlooked in earlier research. **First,** Li *et al.* [24] adopted a language model (*e.g.*, GloVe [38]) pre-trained on solely

[1] In our field, a snippet refers to a set of consecutive frames composed of 16 frames.

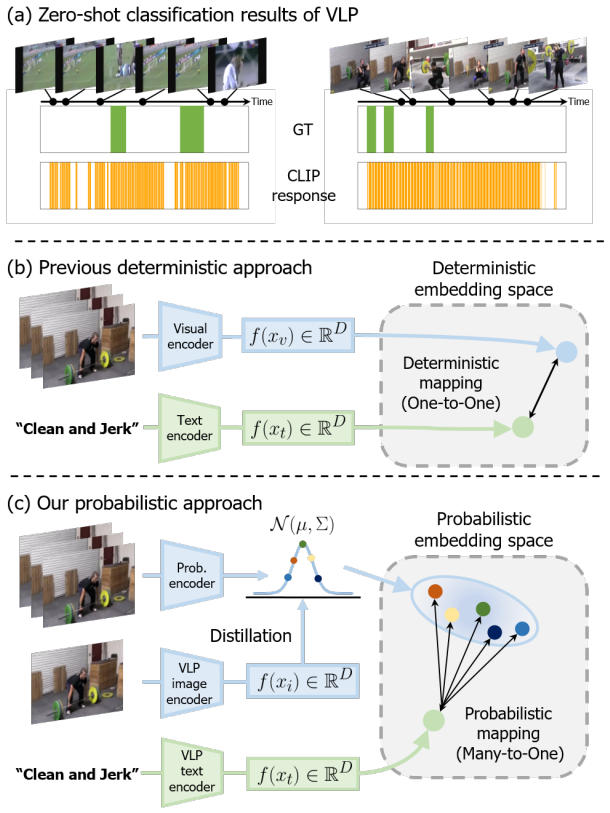

Figure 1: (a) CLIP's deterministic pre-training with image-text pairs fails to equip it with the necessary understanding of fine-grained human motion variations. (b) Earlier studies have primarily focused on the direct mapping between language models and visual input based on deterministic representation. (c) The proposed framework utilizes probabilistic embedding and aligns VLP knowledge.

the text modality, resulting in inadequate initialization for alignment with the human action pre-trained visual feature. **Second,** Chen *et al.* [17] proposed an alternative optimization strategy to introduce an effective distillation framework. However, this alternative optimization scheme requires manually identifying the optimal settings according to the dataset. **Third,** and most importantly, we observe that the utilization of deterministic representation in previous studies for incorporating text information is not suitable for human action understanding.

To confirm this, we conducted an analysis of zero-shot classification with CLIP [39], a prominent study in the domain of VLP. As shown in Figure 1(a), we compared the similarity response between the text prompt (*i.e.*, "a frame of [CLS]") representation and the corresponding frame visual representation. It reveals a high level of activation even when actual human actions do not occur, as long as there is visual relevance to the action text category. This is because CLIP was pre-trained, considering only one-to-one matching between a single image and its caption. The previous research depicted in Figure 1(b) cannot address the aforementioned issue as it solely relies on deterministic representation via one-to-one matching, making it challenging to capture fine-grained human motion. Furthermore, the lack of consideration for direct alignment with pre-trained human action knowledge results in insufficient temporal dynamics modeling.

To overcome this issue, we introduce a novel framework, **PVLR**, **P**robabilistic **V**ision **L**anguage **R**epresentation for Weakly supervised Temporal Action Localization, which integrates VLP knowledge and human action knowledge within the probabilistic embedding space, as shown in Figure 1(c). To begin with, pre-trained human action knowledge, such as Kinetics [2], is utilized to initialize a probabilistic embedding space. In this step, probabilistic adapters are introduced to estimate parameters for the snippet-level probability distribution. Subsequently, we transfer the large-scale VLP knowledge to the estimated probability distribution to create a joint probabilistic embedding space. To capture the temporal dynamics of action, we obtain samples from the estimated probability distribution to offer diverse perspectives, many-to-one matching, then measure their similarity with category text embedding via Monte-Carlo estimation.

Furthermore, to learn a distinctive embedding space, we propose a distribution contrastive learning scheme to capture the statistical similarity between distributions. We enhance intra-class compactness by learning the similarity of content (action or background) within videos and maximize inter-class separability by leveraging action category information across videos. To enhance the intra-class compactness, we draw inspiration from snippet mining in prior work [55] to differentiate a similar snippet distribution among related content. For inter-class separability, we build a video-level probabilistic distribution based on a Gaussian mixture model (GMM) and make the mixture distribution separable between different action classes. To the best of our knowledge, this is the first attempt to investigate multimodal probabilistic representations for weakly supervised temporal action localization. Our main contributions in this work are summarized as follows:

(1) We introduce a novel framework that aligns VLP knowledge and action knowledge within a probabilistic space to fully consider temporal dynamics for fine-grained motion modeling.
(2) We propose an intra- and inter-distribution contrastive strategy based on statistical distance to construct a distinctive probabilistic embedding space.
(3) We conduct extensive experiments and ablation studies to demonstrate the significance of the probabilistic embedding and the proposed method, showing superior performance on two public benchmarks (THUMOS14 and ActivityNet v1.3).

2 RELATED WORK

2.1 Weakly Supervised Temporal Action Localization

Weakly supervised temporal action localization (WTAL) is proposed to alleviate the laborious annotation procedure for Temporal Action Localization, training with only video-level labels. In the early stages of research, Multiple Instance Learning (MIL)-based approaches [10, 14, 21, 22, 33, 37, 45] were proposed, treating a video as a bag of multiple action and background instances. Zhang *et al.* [55] introduced a snippet contrast loss, refining the representation of ambiguous instances in the feature space through

snippet mining and contrastive learning. Subsequently, several approaches [12, 25, 42, 61] generated snippet-level pseudo labels to explicitly guide the model as a localization-by-localization framework. However, pseudo labels generated based on video-level supervision were inaccurate and noisy, making it challenging to achieve the desired performance.

Recently, approaches utilizing the semantic information of action category names have emerged to address the fundamental absence of temporal annotation in WTAL [17, 24]. Li *et al.* [24] designed a novel framework with a discriminative objective to enlarge inter-class differences and a generative objective to enhance intra-class integrity via text information. Chen *et al.* [17] proposed a novel distillation and collaboration framework with complementary Classification Based Pre-training (CBP) and Vision-Language Pre-training (VLP) branches. While these works distinguish themselves with promising performances without additional annotation costs, there is still potential for further development. In our framework, we integrate VLP knowledge and human action knowledge within the probabilistic space previously unexplored in existing literature, enabling a diverse understanding of human action.

2.2 Vision Language Pre-training

Vision language pre-training (VLP) learns a joint representation through large-scale image-text pair datasets with consistent contextual information. A representative work is CLIP [39], mapping image-text pairs with consistent contextual information into the visual and textual encoders separately and facilitating the learning of a joint embedding space through aligned representations. CLIP has shown great success in many image understanding tasks, including image classification [6, 34], semantic segmentation [20, 40], image generation [7, 43], and visual question answering [35]. Building upon the success of CLIP in the image domain, some research efforts [29–31, 52] have emerged that aim to leverage CLIP's vision-language representation in the video domain. Our work is also a contribution to the research aimed at extending VLP knowledge into the realm of untrimmed video and human action understanding.

2.3 Probabilistic Representation

The main idea of probabilistic embedding is to map inputs to probability distributions in the embedding space. Specifically, probabilistic representation has shown a strong potential to embed asymmetric relations, quantify uncertainty, add robustness and more. To achieve this objective, the desired distributions are estimated by a deep neural network and optimized to maximize their likelihood. PCME [5] models one-to-many relationships in the joint embedding space with uncertainty estimation and introduces a soft cross-modal contrastive loss. ProViCo [36] proposed self-supervised video representation learning that bridges contrastive learning with probabilistic embedding with Gaussian mixture model. ProbVLM [48] utilizes a probabilistic adapter that estimates probability distributions for embeddings of a vision-language pre-trained model through inter- and intra-modal alignment in a post-hoc manner, without requiring extensive datasets or intensive computing. The objective of this study is to transfer the knowledge of a pre-trained vision-language model into the probabilistic embedding space, with an explicit objective of enhancing human action understanding.

3 METHOD

3.1 Base Approach

3.1.1 Base Head. Our works will introduce the probabilistic vision-language representation into a basic MIL approach. To begin, we detail the fundamental base head of the MIL approach in this section. We split each video into multi-frame, non-overlapping snippets and sample a fixed number T of snippets to handle variations in video lengths. After sampling, it is common practice to use a pre-trained snippet feature extractor [2] for RGB $\mathbf{X}^R = \{\mathbf{x}_t^r\}_{t=1}^T$ and optical flow $\mathbf{X}^O = \{\mathbf{x}_t^o\}_{t=1}^T$ representation. Afterwards, we concatenate features from each modality $[\mathbf{X}^R; \mathbf{X}^O] \in \mathbb{R}^{T\times 2D}$ and feed them into the base head f_{base}, generating the fused base feature $\mathbf{X}^B \in \mathbb{R}^{T\times 2D}$, expressed as:

$$\mathbf{X}^B = f_{base}([\mathbf{X}^R; \mathbf{X}^O]; \phi_{base}) \in \mathbb{R}^{T\times 2D}, \tag{1}$$

where f_{base} is mainly implemented with a series of temporal convolution with ReLU activation. In addition, an actionness attention weight $\mathbf{a} \in \mathbb{R}^{T\times 1}$ is generated to differentiate between the foreground and the background region:

$$\mathbf{a} = \frac{\mathcal{A}(\mathbf{X}^R, \mathbf{X}^O) + \mathcal{A}(\mathbf{X}^O, \mathbf{X}^R)}{2} \in \mathbb{R}^{T\times 1}, \tag{2}$$

where $\mathcal{A}(\cdot)$ is an attention branch consisting of several temporal convolutional layers. Following the MIL framework, we feed the base feature $\mathbf{X}^B$ into the classification head f_{cls} to generate the base class activation sequence (CAS):

$$\mathbf{S}^{base} = f_{cls}(\mathbf{X}^B; \phi_{cls}) \in \mathbb{R}^{T\times(C+1)}, \tag{3}$$

where C is the number of action classes. To handle background regions in untrimmed videos, we add an auxiliary class to model the background. We then aggregate snippet-level activation scores, to obtain video-level class prediction $\mathbf{p}^{base} = \mathcal{K}(\mathbf{S}^{base}) \in \mathbb{R}^{C+1}$, where $\mathcal{K}(\cdot)$ represents the top-k average pooling along the temporal axis. After obtaining the video-level category prediction, we construct a loss function $\mathcal{L}_{base}$ using cross-entropy loss as follows:

$$\mathcal{L}_{base} = -\sum_{c=1}^{C+1} \mathbf{y}^{base} \log(\mathbf{p}_c^{base}), \tag{4}$$

where $\mathbf{y}^{base} = [y_1, \cdots, y_C, 1] \in \mathbb{R}^{C+1}$ is the video-level label with an auxiliary background class. Background-suppressed CAS can be acquired by applying the attention weight $\mathrm{S}_{supp} = \mathbf{a} \otimes \mathbf{S}^{base}$. We can also build a loss function $\mathcal{L}_{supp}$ with background suppressed video-level class score $\mathbf{p}^{supp} = \mathcal{K}(\mathrm{S}_{supp}) \in \mathbb{R}^{C+1}$ as follows:

$$\mathcal{L}_{supp} = -\sum_{c=1}^{C+1} \mathbf{y}^{supp} \log(\mathbf{p}_c^{supp}), \tag{5}$$

where $\mathbf{y}^{supp} = [y_1, \cdots, y_C, 0] \in \mathbb{R}^{C+1}$ is the video-level label without a background class. Optimizing $\mathcal{L}_{cls} = \mathcal{L}_{base} + \mathcal{L}_{supp}$, enables the model to distinguish snippets that significantly contribute to video-level action classification.

Our proposed method can be applied to any base head based on the MIL approach. We adopt CO_2-Net [10] as the foundation to derive video-level class predictions, chosen for its simplicity and well-documented code. Also, we incorporate the previously

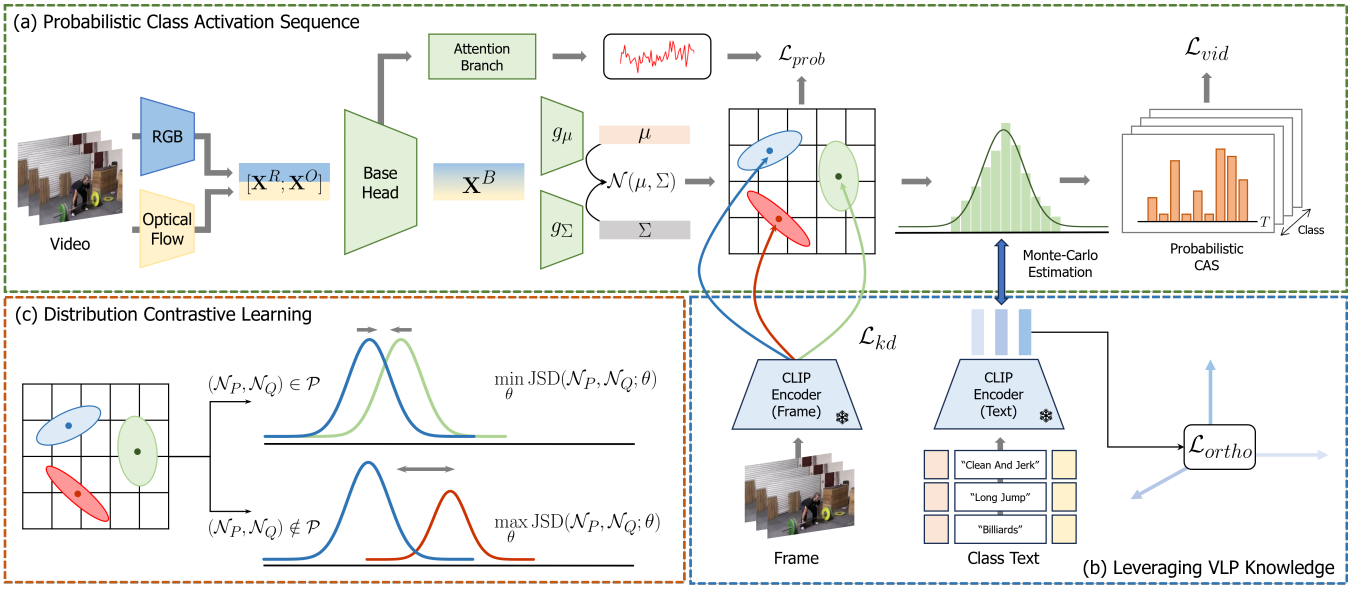

Figure 2: Overview of the proposed PVLR. (a) Probabilistic Class Activation Sequence: For the probabilistic embedding, probabilistic adapters are augmented to facilitate the estimation of probabilistic distributions for individual snippets. (b) Leveraging VLP knowledge: We estimate probabilistic distributions and guide the model with semantic textual information corresponding to action categories. (c) Distribution Contrastive Learning: By training statistical similarities from probabilistic distribution, we aim to build distinctive embedding space.

introduced $\mathcal{L}_{oppo}$, $\mathcal{L}_{norm}$ and $\mathcal{L}_{guide}$ to enhance the optimization of the base head. As these losses were proposed in previous works [14, 21, 22, 32, 37], we do not claim originality for them. The overall objective of the baseline approach $\mathcal{L}_{vid}$ is defined as:

$$\mathcal{L}_{vid} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{oppo} + \lambda_3 \mathcal{L}_{norm} + \lambda_4 \mathcal{L}_{guide}. \tag{6}$$

3.2 Probabilistic Class Activation Sequence

In this section, we reformulate CAS into a probabilistic class activation sequence (P-CAS) to effectively utilize VLP knowledge within a probabilistic embedding space. To achieve this, we model the probabilistic distribution $p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\theta)$ and estimate the parameters θ, optimizing neural networks using human action and VLP knowledge. The full framework is depicted in Figure 2.

3.2.1 Probabilistic Embedding. Initially, we establish probabilistic embedding space by leveraging pre-trained human action knowledge on Kinetics [2]. From the base feature $\mathbf{X}^B = \{\mathbf{x}_t\}_{t=1}^T \in \mathbb{R}^{T\times 2D}$, we formulate a snippet-level probability distribution $p(\mathbf{z}|\mathbf{x}_t)$ as a multivariate Gaussian distribution with a mean vector and a diagonal covariance matrix to model the probabilistic embedding space:

$$p(\mathbf{z}|\mathbf{x}_t) \approx \mathcal{N}(g_\mu(\mathbf{x}_t), \mathrm{diag}(g_\Sigma(\mathbf{x}_t))), \tag{7}$$

where g_μ is an embedding layer that estimates the mean vector $g_\mu(\mathbf{x}_t) \in \mathbb{R}^D$ and g_Σ is an embedding layer that estimates covariance matrix $g_\Sigma(\mathbf{x}_t) \in \mathbb{R}^D$ of the target Gaussian. With the estimated $p(\mathbf{z}|\mathbf{x}_t)$, we can sample K random embeddings $\mathbf{z}^{(k)} \in \mathbb{R}^D$ that can represent the estimated distribution following [19]:

$$\mathbf{z}_t^{(k)} = g_\mu(\mathbf{x}_t) + \epsilon^{(k)} \cdot g_\Sigma(\mathbf{x}_t) \in \mathbb{R}^D, \tag{8}$$

where $\epsilon^{(k)} \in \mathbb{R}^D$ are independently and identically sampled from a D-dimensional unit Gaussian. Our goal is to utilize K embeddings sampled from the estimated probability distribution for each snippet to capture human actions from a more diverse range of perspectives. Additionally, for textual information, we transform action category names into pre-trained embeddings. We can achieve this by freezing the CLIP text transformer $\Psi_C(\cdot)$ and extracting the embeddings $\mathbf{X}^C = \{\mathbf{x}_c\}_{c=1}^{C+1}$:

$$\mathbf{x}_c = \Psi_C([\mathbf{L}_s; \Psi_{emb}(t_c); \mathbf{L}_e]) \in \mathbb{R}^D, \tag{9}$$

where $\mathbf{L}_s, \mathbf{L}_e$ are learnable tokens, t_c refers to action category, and Ψ_{emb} is word embedding layer. P-CAS is then defined by determining the action confidence score along the temporal axis with the estimated probability distribution and action category representation. Specifically, to measure the confidence score between the estimated distribution and category representation, we formulate P-CAS as $\mathbf{S}_{prob} \in \mathbb{R}^{T\times(C+1)}$ via Monte-Carlo estimation:

$$s_{prob}(t,c) \approx \frac{1}{K}\sum_{k=1}^{K} \mathrm{sim}(\mathbf{z}_t^{(k)}, \mathbf{x}_c)/\tau, \tag{10}$$

where $\mathrm{sim}(\cdot)$ means cosine similarity and τ is temperature parameter. Besides, to enhance the differentiation among action categories, we design an orthogonal loss $\mathcal{L}_{ortho}$ to ensure the uniqueness of each category representation:

$$\mathcal{L}_{ortho} = \left\| \mathbf{X}^C (\mathbf{X}^C)^\mathrm{T} - \mathbf{I} \right\|_\mathrm{F}^2, \tag{11}$$

where $\mathbf{I}$ is the identity matrix and $\|\cdot\|_\mathrm{F}^2$ is the Frobenius norm of a matrix.

3.2.2 VLP Knowledge Distillation. However, during the estimation of the current probability distribution, only the textual information from VLP is utilized, overlooking the alignment between human action knowledge and the visual representation provided by VLP. Therefore, we aim to integrate VLP visual knowledge into the probability distribution estimation process.

Estimating the entire distribution can be challenging due to the deterministic pre-training of CLIP. However, large-scale pre-trained representations can offer a generalized point approximation (*e.g.*, mean vector) for the desired distribution. To achieve this, our probabilistic embedding utilizes the CLIP's deterministic representations as estimates for the mean, $g_\mu(\mathbf{x}_t)$ of the targeted distribution. To transfer VLP knowledge into the probabilistic embedding space, we first sample a set of frames $\{f_t\}_{t=1}^{T}$ with a fixed temporal stride from the video. Next, we freeze the CLIP Image Encoder $\Psi_{\mathcal{I}}(\cdot)$ and extract the embeddings $\mathbf{X}^{\mathcal{I}} = \{\Psi_{\mathcal{I}}(f_t)\}_{t=1}^{T} \in \mathbb{R}^{T\times D}$. For a pair of snippet feature and CLIP image feature $(\mathbf{x}_t^b, \mathbf{x}_t^i)$, the distillation loss $\mathcal{L}_{kd}$ is defined as:

$$\mathcal{L}_{kd} = -\frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{2}\left(\frac{g_\mu(\mathbf{x}_t)\cdot\mathbf{x}_t^i}{\|g_\mu(\mathbf{x}_t)\|\,\|\mathbf{x}_t^i\|}+1\right)\right). \tag{12}$$

We utilize the rescaled cosine similarity between the estimated mean $g_\mu(\mathbf{x}_t^b)$ and CLIP image representation $\mathbf{x}_t^i$ as the matching score. The objective of $\mathcal{L}_{kd}$ is to align the estimated mean $g_\mu(\mathbf{x}_t^b)$ with the generalized fixed point $\mathbf{x}_t^i$ of CLIP embedding, thus transferring pre-trained CLIP knowledge into the desired probability distribution.

3.3 Distribution Contrastive Learning

We formulate a probabilistic embedding space by aligning human action knowledge with VLP knowledge. However the crucial factor of distributional similarity remains unexplored. Distributions corresponding to human actions should exhibit similarities with each other while contrasting with background distributions. To address this, we aim to enhance the completeness of the probabilistic embedding space through distribution contrastive learning based on statistical distances.

3.3.1 Intra-Distribution Contrastive Learning. We begin by considering contrastive learning between distributions within the video. Action distributions within a video are expected to share similarities while remaining distinct from the background distributions. To achieve these objectives, we adopt the snippet mining algorithm of CoLA [55], which uses attention weight $\mathbf{a} \in \mathbb{R}^{T\times 1}$ to differentiate between action and background snippets within the video. We classify easily distinguishable samples as easy actions (top-k) and easy backgrounds (bottom-k) based on the actionness score. However, boundary-adjacent snippets are less reliable due to their transitional position, making detection hard. To mine hard action and background snippets, we threshold the attention weight (1 indicates action, 0 indicates background):

$$\mathbf{b}^{(t)} = \begin{cases} 1 & \text{if} \quad \mathbf{a}^{(t)} > \theta_b \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

where θ_b is the threshold value.

We utilize the same strategy performing, cascaded dilation or erosion operations to identify challenging samples (hard to differentiate) at the action/background boundaries.

$$\mathcal{R}_{inner} = (\mathbf{b}; m)^- - (\mathbf{b}; \mathcal{M})^- \tag{14}$$

$$\mathcal{R}_{outer} = (\mathbf{b}; \mathcal{M})^+ - (\mathbf{b}; m)^+, \tag{15}$$

where $(\cdot)^-$ and $(\cdot)^+$ represent the binary erosion and dilation operations with smaller mask m and larger mask $\mathcal{M}$. Following earlier work [55], we consider inner regions as hard action snippet sets and outer regions as hard background snippet sets. Here, we define hard actions as having a positive relation $\mathcal{P}_{act}$ with easy actions (top-k) and hard backgrounds as having a positive relation $\mathcal{P}_{bkg}$ with easy backgrounds (bottom-k). After snippet mining, we utilize KL divergence as a statistical metric to measure the similarity of snippet distributions. The KL divergence between multivariate Gaussian is defined as:

$$\begin{aligned}\mathrm{KL}(\mathcal{N}_P \parallel \mathcal{N}_Q) = \frac{1}{2}(&\mathrm{tr}(\Sigma_Q^{-1}\Sigma_P)+\\ &(\mu_Q-\mu_P)^{\mathrm{T}}\Sigma_Q^{-1}(\mu_Q-\mu_P)+\ln\left(\frac{\det\Sigma_Q}{\det\Sigma_P}\right)-D).\end{aligned} \tag{16}$$

To ensure intra-class compactness of the embedding space, we propose an intra-contrastive loss $\mathcal{L}_{intra}$ to refine snippet-level distribution similarity. The intra-contrastive loss $\mathcal{L}_{intra}$ is formulated as:

$$\mathcal{L}_{intra} = \begin{cases} -\log(1-p(\mathcal{N})) & \text{if } (\mathcal{N}_P, \mathcal{N}_Q) \in \mathcal{P} \\ -\log(p(\mathcal{N})) & \text{otherwise} \end{cases}, \tag{17}$$

where $\mathcal{P}$ represents positive sets for intra-distribution contrastive learning and $(\mathcal{N}_p, \mathcal{N}_Q)$ represents two arbitrary Gaussians. We formulate the matching probability $p(\mathcal{N})$ as the Jensen-Shannon divergence $\mathrm{JSD}(\mathcal{N}_P, \mathcal{N}_Q)$, based on KL divergence.

3.3.2 Inter-Distribution Contrastive Learning. We further introduce inter-distribution contrastive learning utilizing action category labels to ensure inter-class separability. Here, we represent the whole video distribution $p(\mathbf{z}|V)$ as a Gaussian mixture model (GMM) to measure video-level similarity,

$$p(\mathbf{z}|V) \approx \sum_{t=1}^{T}\mathbf{a}_t\cdot\mathcal{N}(g_\mu(\mathbf{x}_t), \mathrm{diag}(g_\Sigma(\mathbf{x}_t))). \tag{18}$$

To estimate $p(\mathbf{z}|V)$, we use the attention weight $\mathbf{a} \in \mathbb{R}^T$ as a mixing coefficient to appropriately combine distributions based on the actionness score. Given video-level category labels, we formulate a self-similarity map $\mathbf{H} \in \mathbb{R}^{N\times N}$ (1 for the same class, 0 for different) to characterize relationships between videos. Similar to intra-contrastive learning, we compute the matching probabilities $p(\mathcal{N})$ between N videos within a mini-batch across mixture models, and enhance inter-video representation by comparing them with the self-similarity map. Finally, the inter-contrastive loss $\mathcal{L}_{inter}$ is formulated as:

$$\mathcal{L}_{inter} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{L}_{\mathrm{BCE}}(\mathbf{H}(i,j), p(\mathcal{N})), \tag{19}$$

where $\mathcal{L}_{\mathrm{BCE}}$ is a binary cross entroy loss. Beyond aligning with VLP knowledge and the probabilistic embedding space, our proposed contrastive learning framework enforces constraints on the

Table 1: Comparison with previous state-of-the-art methods on THUMOS14. 0.1:0.7 and 0.1:0.5 represent the average mAP under IoU thresholds of 0.1:0.7 and 0.1:0.5.

Supervision	Method	Venue	mAP@IoU (%)							AVG	
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.7	0.1:0.5
Fully supervised	TAL-Net [3]	CVPR 2018	59.8	57.1	53.2	48.5	42.8	33.8	20.8	45.1	52.3
	P-GCN [54]	CVPR 2019	69.5	67.8	63.6	57.8	49.1	-	-	-	61.6
	BUMR [59]	ECCV 2020	-	-	53.9	50.7	45.4	38.0	28.5	-	-
Weakly supervised	CoLA [55]	CVPR 2021	66.2	59.5	51.5	41.9	32.2	22.0	13.1	40.9	50.3
	CO_2-Net [10]	MM 2021	70.1	63.6	54.5	45.7	38.3	26.4	13.4	44.6	54.4
	Xu *et al.* [51]	TPAMI 2023	73.1	66.9	58.3	48.8	36.5	24.4	13.4	45.9	56.7
	Li *et al.* [24]	CVPR 2023	71.1	65.0	56.2	47.8	39.3	27.5	15.2	46.0	55.9
	Wang *et al.* [50]	CVPR 2023	73.0	68.2	60.0	47.9	37.1	24.4	12.7	46.2	57.2
	P-MIL [41]	CVPR 2023	71.8	67.5	58.9	49.0	40.0	27.1	15.1	47.0	57.4
	AHLM [49]	ICCV 2023	**75.1**	68.9	60.2	48.9	38.3	26.8	14.7	47.2	58.3
	DDG-Net [46]	ICCV 2023	72.5	67.7	58.2	49.0	41.4	27.6	14.8	47.3	57.8
	STCL-Net [8]	TPAMI 2023	72.7	67.1	58.2	49.7	41.8	28.7	16.0	47.7	57.9
	GauFuse [61]	CVPR 2023	74.0	69.4	60.7	51.8	42.7	26.2	13.1	48.3	59.7
	Chen *et al.* [17]	CVPR 2023	73.5	68.8	**61.5**	**53.8**	42.0	29.4	16.8	49.4	60.0
	ISSF [53]	AAAI 2024	72.4	66.9	58.4	49.7	41.8	25.5	12.8	46.8	57.8
	PVLR (Ours)	MM 2024	74.9	**69.9**	61.4	53.1	**45.1**	**30.5**	**17.1**	**50.3**	**60.9**

probabilistic representation to ensure both intra-compactness and inter-separability.

3.4 Total Objectives

Considering all the previously mentioned objectives, the total objective $\mathcal{L}_{total}$ of the entire framework is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{vid} + \alpha\mathcal{L}_{kd} + \beta\mathcal{L}_{ortho} + \gamma(\mathcal{L}_{intra} + \mathcal{L}_{inter}), \quad (20)$$

where α, β, γ are hyper-parameters to balance these loss terms. We kept the parameters related to $\mathcal{L}_{vid}$ unchanged and performed a grid search only on α, β and γ. During testing, we use the same strategy as previous research [10, 24] to extract proposal candidates for an input video. Lastly, we apply soft non-maximum suppression to eliminate overlapping proposals.

4 EXPERIMENTS

4.1 Experimental Settings

We conduct experiments on two popular WTAL benchmarks: THUMOS14 [13] and ActivityNet v1.3 [1]. THUMOS14 is a widely used benchmark for the WTAL problem, containing 200 validation videos and 213 test videos across 20 sports categories. Following previous works [10, 42, 55], we use the 200 validation videos for training our framework and the 213 test videos for evaluation. THUMOS14 is the most challenging dataset in WTAL due to motion blur, significant intra-class variations, and extremely short action instances. ActivityNet v1.3 has 10,024 training videos, 4,926 validation videos, and 5,044 testing videos from 200 action categories. Since the testing set annotations are not released, we train on the training set and test on the validation set. Challenges in ActivityNet typically involve the large number of action categories. Following standard evaluation metrics, we evaluate our method using mean Average Precision (mAP) under different Intersection over Union (IoU) thresholds on the temporal axis.

4.2 Implementation Details

To conduct experiments on the THUMOS14 dataset and ActivityNet v1.3 dataset, we first divide each video into non-overlapping segments consisting of 16 frames. Subsequently, we extract the 1024-dimensional RGB and optical flow features from the I3D network [2] pre-trained on the Kinetics400 dataset. We apply the TV-L1 algorithm to extract the flow features. The fixed number of segments T is set to 320 for THUMOS14 and 60 for ActivityNet v1.3. We adopt ResNet-50 as a backbone network for the CLIP image encoder Ψ_I. It is worth noting that the I3D network and the CLIP encoders are not fine-tuned during training. For CLIP image feature, we divide the video as described above and the middle frame of each snippet is fed into Ψ_I. For the probabilistic adapter, g_μ indicates a single linear layer, while g_Σ is a separate network with a linear layer followed by the ReLU function, to ensure the Σ remains positive definite. Similar to previous work [16], we prepend and append 4 prompt vectors to word embedding $\Psi_{emb}(t_c)$, which is initialized with $\mathcal{N}(0, 0.01)$. In P-CAS, we use the learnable random embedding to model the background class, which is hard to characterize. Our experiments are conducted on an NVIDIA Tesla V100 GPU.

4.3 Comparison With State-Of-The-Art Methods

In this section, we compare our proposed PVLR with previous state-of-the-art methods. For THUMOS14, it is evident that the proposed PVLR outperforms all previous state-of-the-art methods, as shown in Table 1. Notably, in WTAL scenarios where performance under high IoU (0.5-0.7) is crucial, our method surpasses all existing methodologies. In direct comparison to prior studies [17, 24] that

Table 2: Results on ActivityNet v1.3. 0.5:0.95 indicates the average mAP at IoU thresholds of 0.5:0.95.

Method	Venue	mAP@IoU (%)			AVG
		0.5	0.75	0.95	0.5:0.95
DCC [25]	CVPR 2022	38.8	24.2	5.7	24.3
RSKP [12]	CVPR 2022	40.6	24.6	5.9	25.0
ASM-Loc [9]	CVPR 2022	41.0	24.9	6.2	25.1
STCL-Net [8]	TPAMI 2023	40.6	24.0	6.0	24.7
Zhang *et al.* [58]	TCSVT 2023	41.6	25.1	6.5	25.3
LPR [11]	TCSVT 2023	41.4	25.3	6.2	25.4
P-MIL [41]	CVPR 2023	41.8	25.4	5.2	25.5
AHLM [49]	ICCV 2023	42.3	24.8	**6.9**	25.9
Li *et al.* [24]	CVPR 2023	41.8	26.0	6.0	26.0
Wang *et al.* [50]	CVPR 2023	41.8	25.7	6.5	26.3
Li *et al.* [23]	TNNLS 2023	42.3	26.4	6.1	26.4
CASE [29]	ICCV 2023	43.2	26.2	6.7	26.8
Yun *et al.* [53]	AAAI 2024	39.4	25.8	6.4	25.8
SRHN [60]	TCSVT 2024	41.7	26.1	6.1	26.2
Liu *et al.* [28]	ICASSP 2024	42.8	26.8	6.0	26.4
PVLR (Ours)	MM 2024	**43.6**	**27.4**	6.5	**27.4**

Table 3: Component-wise ablation study on THUMOS14.

Method	mAP@IoU			AVG
	0.3	0.5	0.7	0.3:0.7
Baseline	53.0	36.5	13.6	34.4
+Distillation from VLP knowledge	58.1	40.3	15.3	38.1
+Intra-contrastive	59.4	42.3	15.9	39.5
+Inter-contrastive	59.6	43.7	16.9	40.1
+Orthogonalization of text prompts	**61.4**	**45.1**	**17.1**	**41.4**

Table 4: Further analysis for probabilistic representation.

Metric	mAP@IoU			AVG
	0.3	0.5	0.7	0.3:0.7
Deterministic CAS	57.5	41.0	15.8	38.3
Mahalanonis Distance	60.1	44.0	17.3	40.7
Bhattacharyya Distance	60.8	44.0	17.3	40.9
Kullback–Leibler Divergence	**61.4**	**45.1**	**17.1**	**41.4**

also incorporate textual information, our approach demonstrates superior performance, with a margin ranging from 0.9% to 4.3% in average mAP (0.1:0.7). Additionally, our approach either outperforms or reaches similar performance levels as recent fully supervised methods. In Table 2, results for the larger dataset ActivityNet v1.3 are presented. Similarly, our proposed PVLR shows superior performance compared to existing weakly supervised state-of-the-art methods.

4.4 Ablation Study

To demonstrate the effectiveness of our model components, we analyze the impact of each component in this section with THUMOS14. In Table 3, the baseline is reported based on only $\mathcal{L}_{vid}$ without probabilistic embedding. The VLP knowledge distillation module serves as a pivotal step within our framework, marking the inception of our approach. By conducting feature alignment in a probabilistic space, PVLR introduces a fundamental basis that was previously overlooked in earlier literature. Integrating VLP knowledge results in a performance boost of 3.7% through the implementation of a probabilistic class activation sequence (P-CAS). Additionally, refining our probabilistic embedding space with distribution contrastive learning leads to a 2.0% improvement. Finally, introducing orthogonalization of text embeddings enhances the discriminative capacity between text category embeddings, yielding a 1.3% gain. We ultimately demonstrate the effectiveness of our proposed module by achieving a performance improvement of 7.0%, a level that is difficult to find in previous research.

4.5 Discussion

To provide deeper insights into the design aspect of our proposed framework, we conduct several experiments in this section.

4.5.1 Probabilistic Representation. As the initial procedure in our framework, probabilistic representation is of great importance. To validate its significance, we develop a simple baseline using deterministic representation. For the deterministic baseline, we conduct experiments utilizing one-to-one matching between human action knowledge and text embedding without probabilistic adapter. In Table 4, the first row "Deterministic CAS" indicates the deterministic baseline. Quantitatively, there is a performance decrease of about 3.1% compared to the proposed probabilistic approach in Table 4. We also compare qualitative visualization results of selected videos from the THUMOS14 dataset. Figure 3 illustrates that the deterministic approach frequently produces predictions beyond the ground truth boundaries, struggling to capture subtle variations in human action. In contrast, the probabilistic method effectively models temporal dynamics, focusing its predictions on the segments where real actions unfold. In Figure 3(a), the probabilistic approach appears to struggle with completely filling the GT segment, yet this specific area, characterized by an absence of motion change, is designated for future exploration. From Table 4, besides the KL divergence, other statistical metrics are also suitable for our contrastive learning. Table 4 reveals that metrics capable of assessing inter-distributional similarity exhibit relatively consistent performance with minimal variation. By not relying on a specific distance metric, it can be considered that a well-generalized probability distribution has been estimated, leading to the successful modeling of a probabilistic embedding space. Finally, the marginal superiority of KL divergence leads to its utilization for the proposed distribution contrastive learning.

4.5.2 Computational Cost. As shown in Table 6, we measure the model complexity in terms of additional text features (Feature) and multiply-accumulative operations (MACs), the number of trainable parameters (Params), and running time (Time). Since we only utilize a lightweight probabilistic encoder (a single linear layer) with our

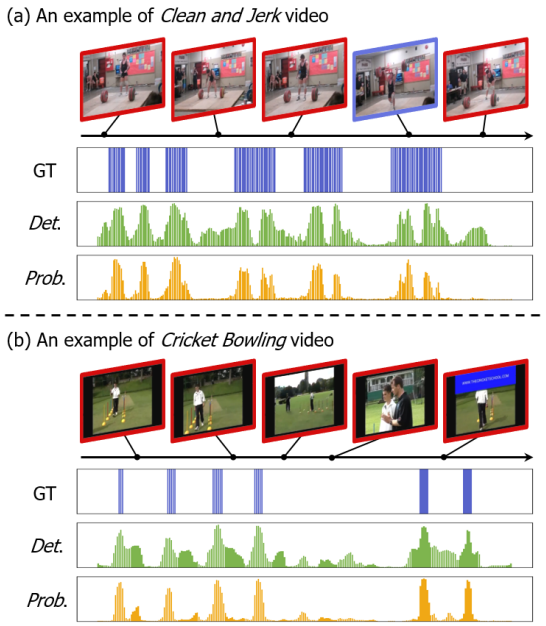

Figure 3: Qualitative Results on THUMOS14. We compared the class activation sequence (CAS) of deterministic and probabilistic approaches. In this case, the red box is for the background, and the blue box is for the action.

base approach, CO_2-Net [10], it shows nearly similar complexity. Li's approach [24], which includes GloVe [38] embeddings, shows higher complexity as a result of the dual branch optimization. The official source code for Chen's approach [17], which also utilizes CLIP like ours, is unavailable; thus, its complexity cannot be determined. Even though our work does not claim efficiency as its main contribution, it still shows a competitive trade-off, achieving the best performance.

4.5.3 Number of K samples. To analyze the impact of the number of samples during generation of P-CAS, we compare the performances under different values of K, as shown in Table 5. As observed, a small value for K leads to suboptimal performance, resulting in a lack of representation of the estimated distribution. Here, we denote the previously described deterministic baseline as $K = 0$. Considering that larger values of K capture the entire distribution of the snippet through Monte-Carlo estimation, performance improves with an increase in K. Nevertheless, an increased value for K results in higher computational demands. Calculating the confidence score for composing a P-CAS requires computations on the order of $O(K)$ for each snippet and action category. Considering the computational overhead, we decide on $K = 20$.

4.6 Generalization Study

In Table 7, we demonstrate the generality of our contributions by integrating them into previous works in a plug-and-play manner. To achieve this, we conduct comparative experiments by replacing the base WTAL head with those from previous works [21, 55]. The additional training modules exclusively consider the proposed probabilistic adapter for probabilistic embedding. Furthermore, we reformulate the classification objective using probabilistic class activation sequences (P-CAS). The results show that our framework enhances performance, with an average mAP increase ranging from 2% to 3%, indicating robust generalization across various methods and model architecture designs.

Table 5: Number of K ablation study on THUMOS14.

# of samples	mAP@IoU			AVG
	0.3	0.5	0.7	0.3:0.7
Baseline	53.0	36.5	13.6	34.4
$K = 0$	57.5	41.0	15.8	38.3
$K = 5$	59.8	41.4	15.5	38.9
$K = 10$	60.1	43.8	16.9	40.5
$K = 20$	**61.4**	**45.1**	**17.1**	**41.4**

Table 6: Computational cost comparison on THUMOS14.

	Feature	Params	MACs	Time	AVG
CO_2-Net [10]	-	34.1M	20.9G	1.14s	44.6
Li *et al.* [24]	GloVe [38]	96.6M	41.3G	2.20s	46.0
Chen *et al.* [17]	CLIP [39]	-	-	-	49.4
PVLR	CLIP [39]	49.9M	30.3G	1.45s	**50.3**

Table 7: Framework generalization results on THUMOS14.

Method	mAP@IoU			AVG
	0.3	0.5	0.7	0.3:0.7
BaS-Net [21]	44.6	26.6	10.0	27.0
BaS-Net+Ours	50.1	29.2	10.7	**30.2**
CoLA [55]	51.8	34.0	12.5	32.9
CoLA+Ours	56.2	35.5	13.3	**35.1**

5 CONCLUSION AND FUTURE WORKS

In this work, we present a novel framework that leverages a large-scale pre-trained vision-language model to address WTAL. Our motivation arose from the shortcomings of VLP's deterministic representation and the lack of joint alignment consideration in understanding human actions. To address these concerns, we introduce a probabilistic embedding framework aligned with human action and VLP knowledge, enhanced by distribution contrastive learning. Our method significantly outperforms previous approaches on two prominent datasets, revealing the efficacy of probabilistic embedding within the VLP representation. However, the exploration of probabilistic embedding for text data solely represented by action category names remains unexplored. For future work, we will explore leveraging the recently acclaimed large-language model (LLM) to generate attributes for each action category and subsequently integrate them with probabilistic embeddings.

ACKNOWLEDGMENTS

This work was supported by the Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) under Grant 2022-00156345 (ICT Challenge and Advanced Network of HRD, 50%), Grant 2023-00254529 (metaverse support program to nurture the best talents, 25%) and Grant 2020-0-00011 (Video Coding for Machine, 25%)

REFERENCES

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*.

[2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[4] Feng Cheng and Gedas Bertasius. 2022. Tallformer: Temporal action localization with a long-memory transformer. In *Proceedings of the European Conference on Computer Vision*.

[5] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[6] Marcos V Conde and Kerem Turgutlu. 2021. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[7] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Proceedings of the European Conference on Computer Vision*.

[8] Jie Fu, Junyu Gao, and Changsheng Xu. 2023. Semantic and Temporal Contextual Correlation Learning for Weakly-Supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[9] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. 2022. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[10] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. 2021. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*.

[11] Yufan Hu, Jie Fu, Mengyuan Chen, Junyu Gao, Jianfeng Dong, Bin Fan, and Hongmin Liu. 2023. Learning Proposal-aware Re-ranking for Weakly-supervised Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[12] Linjiang Huang, Liang Wang, and Hongsheng Li. 2022. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[13] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding* (2017).

[14] Ashraful Islam, Chengjiang Long, and Richard Radke. 2021. A hybrid attention mechanism for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[15] Won Jo, Geuntaek Lim, Gwangjin Lee, Hyunwoo Kim, Byungsoo Ko, and Yukyung Choi. 2023. VVS: Video-to-Video Retrieval with Irrelevant Frame Suppression. (2023).

[16] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*.

[17] Chen Ju, Kunhao Zheng, Jinxiang Liu, Peisen Zhao, Ya Zhang, Jianlong Chang, Qi Tian, and Yanfeng Wang. 2023. Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[18] Jinah Kim and Jungchan Cho. 2022. Background-aware robust context learning for weakly-supervised temporal action localization. *IEEE Access* (2022).

[19] Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems* (2015).

[20] Hyeongjun Kwon, Taeyong Song, Somi Jeong, Jin Kim, Jinhyun Jang, and Kwanghoon Sohn. 2023. Probabilistic Prompt Learning for Dense Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[21] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. 2020. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[22] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. 2021. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[23] Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Jie Li, and Xinbo Gao. 2023. Weakly Supervised Temporal Action Localization With Bidirectional Semantic Consistency Constraint. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[24] Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. 2023. Boosting Weakly-Supervised Temporal Action Localization with Text Information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[25] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. 2022. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[26] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[27] Jichao Liu, Chuanxu Wang, and Yun Liu. 2019. A novel method for temporal action localization and recognition in untrimmed video based on time series segmentation. *IEEE Access* (2019).

[28] Peng Liu, Chuanxu Wang, and Min Zhao. 2024. Modal Consensus and Contextual Separation for Weakly Supervised Temporal Action Localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[29] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. 2023. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[30] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* (2022).

[31] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*.

[32] Kyle Min and Jason J Corso. 2020. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *Proceedings of the European Conference on Computer Vision*.

[33] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. 2018. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[34] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. 2023. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*.

[35] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2023. CLIP-Guided Vision-Language Pre-training for Question Answering in 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[36] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. 2022. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[37] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision*.

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

[40] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[41] Huan Ren, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. 2023. Proposal-Based Multiple Instance Learning for Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[42] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. 2023. PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* (2022).

[44] Zhengyang Shen, Feng Wang, and Jin Dai. 2020. Weakly supervised temporal action localization by multi-stage fusion network. *IEEE Access* (2020).

[45] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. 2018. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision.*

[46] Xiaojun Tang, Junsong Fan, Chuanchen Luo, Zhaoxiang Zhang, Man Zhang, and Zongyuan Yang. 2023. DDG-Net: Discriminability-Driven Graph Network for Weakly-supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[47] Praveen Tirupattur, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. Modeling multi-label action dependencies for temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[48] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2023. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[49] Guiqin Wang, Peng Zhao, Cong Zhao, Shusen Yang, Jie Cheng, Luziwei Leng, Jianxing Liao, and Qinghai Guo. 2023. Weakly-Supervised Action Localization by Hierarchically-structured Latent Attention Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[50] Yu Wang, Yadong Li, and Hongbin Wang. 2023. Two-Stream Networks for Weakly-Supervised Temporal Action Localization With Semantic-Aware Mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[51] Zhe Xu, Kun Wei, Erkun Yang, Cheng Deng, and Wei Liu. 2023. Bilateral Relation Distillation for Weakly Supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[52] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2023. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *International Conference on Learning Representations* (2023).

[53] Wulian Yun, Mengshi Qi, Chuanming Wang, and Huadong Ma. 2024. Weakly-Supervised Temporal Action Localization by Inferring Salient Snippet-Feature. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[54] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[55] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. 2021. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[56] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision.*

[57] Min Zhang, Haiyang Hu, and Zhongjin Li. 2022. Temporal action localization with coarse-to-fine network. *IEEE Access* (2022).

[58] Songchun Zhang and Chunhui Zhao. 2023. Cross-Video Contextual Knowledge Exploration and Exploitation for Ambiguity Reduction in Weakly Supervised Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[59] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. 2020. Bottom-up temporal action localization with mutual regularization. In *Proceedings of the European Conference on Computer Vision.*

[60] Yibo Zhao, Hua Zhang, Zan Gao, Weili Guan, Meng Wang, and Shengyong Chen. 2024. A Snippets Relation and Hard-Snippets Mask Network for Weakly-Supervised Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[61] Jingqiu Zhou, Linjiang Huang, Liang Wang, Si Liu, and Hongsheng Li. 2023. Improving Weakly Supervised Temporal Action Localization by Bridging Train-Test Gap in Pseudo Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*