

Breaking Temporal Ambiguity: Linguistic Temporal Knowledge Transfer for Weakly-Supervised Action Localization

Zhilin Li¹, Shining Wu¹, Yixin Zhang², Chuandong Cheng^{1,3}, Zilei Wang^{1*}

¹University of Science and Technology of China

²Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

³Department of Neurosurgery,

Centre for Leading Medicine and Advanced Technologies of IHM,

The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, USTC

{lizhilin, wushining}@mail.ustc.edu.cn, {zhyx12, doctorcd, zlwang}@ustc.edu.cn

Abstract

Weakly-supervised temporal action localization (WS-TAL) remains challenging due to the severe temporal ambiguity inherent in video-level supervision. Existing methods often overfit to highly discriminative snippets, struggling to capture the complete action extent. In this paper, we reconceptualize actions not as isolated events, but as multi-stage action evolutions. To alleviate temporal ambiguity, we propose a novel framework driven by the Linguistic Temporal Knowledge Transfer (LTKT) Network, comprising three cohesive modules. First, Temporal Stage Grounding (TSG) leverages explicitly deconstructed LLM descriptions as cross-modal priors, guiding unconstrained visual snippets to align with their corresponding evolutionary stages. Second, Temporal Transition Gating (TTG) enforces forward causality on the timeline. By mandating that an execution stage is flanked by historical preparation and future completion, TTG strictly filters static noise to yield refined temporal action scores. Finally, Linguistic Knowledge Transfer (LKT) integrates knowledge distillation and contrastive learning to embed this structural knowledge into the visual backbone. Extensive experiments on THUMOS14, ActivityNet1.2 and ActivityNet1.3 demonstrate that our method achieves state-of-the-art performance.

1 Introduction

Weakly-supervised temporal action localization (WS-TAL) has garnered significant attention in the computer vision community, as it mitigates the prohibitive cost of frame-level annotations by relying solely on video-level categorical labels. The mainstream paradigm for WS-TAL formulates the task as a Multiple Instance Learning (MIL) problem (Li et al., 2024a; Yun et al., 2024). While these methods have achieved commendable progress in

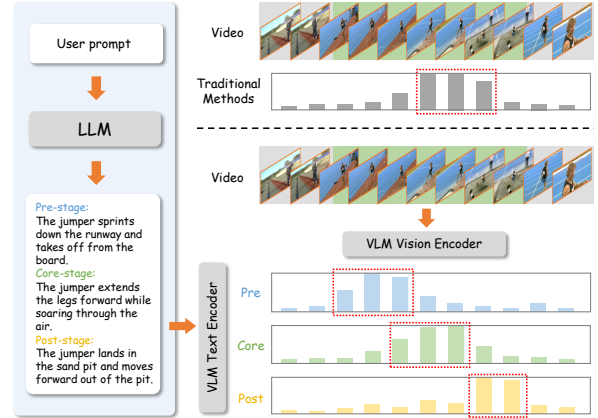


Figure 1: Concept of our Linguistic Temporal Knowledge Transfer. Unlike traditional WS-TAL methods that suffer from temporal ambiguity by overfitting to isolated discriminative snippets (top right), our approach reconceptualizes actions as multi-stage action evolutions. By transferring structural priors from LLMs (left), we anchor visual features into Pre, Core, and Post stages, enforcing temporal causality to achieve highly accurate and complete boundary predictions (bottom right). The green background areas in the video represent the ground truth.

identifying the presence of actions, the inherent absence of fine-grained boundary supervision inevitably plunges the models into severe temporal ambiguity. Without explicit temporal constraints, the network fundamentally lacks the capability to precisely localize when an action starts and when it ends.

Driven predominantly by the classification objective, existing MIL paradigms tend to overfit to highly discriminative snippets. For instance, in a “Long Jump” video, models typically produce anomalously high activations exclusively during the “mid-air” pose, entirely ignoring the preceding run-up and the subsequent landing. Furthermore, these unconstrained models frequently suffer from false positive activations, mistaking static but visually similar background contexts (*e.g.*, an athlete

*Corresponding author: zlwang@ustc.edu.cn

standing still before the jump) for the action itself. Consequently, they struggle to capture the complete action extent and often yield fragmented or contextually illogical boundary estimations. We argue that this limitation stems from treating videos as a bag of isolated frames, thereby completely discarding the temporal relations and the inherent causality of human actions.

To break this temporal ambiguity, we reconceptualize actions not as isolated visual events, but as continuous, multi-stage action evolutions comprising preparation, execution, and completion phases (as illustrated in Figure 1). While this logical progression is challenging to learn purely from unconstrained visual pixels under weak supervision, we observe that Large Language Models (LLMs) encapsulate rich, zero-shot human priors regarding these physical evolutions. We hypothesize that harnessing such linguistic structural knowledge and transferring it into the visual domain is the key to resolving the boundary incompleteness and temporal noise in WS-TAL.

Motivated by this insight, we propose the Linguistic Temporal Knowledge Transfer (LTKT) Network, a novel framework that fundamentally alleviates temporal ambiguity via three cohesive modules. First, the **Temporal Stage Grounding (TSG)** module leverages multi-stage descriptions explicitly deconstructed by LLMs as cross-modal structural priors. Through a global cross-attention mechanism, TSG guides unconstrained visual snippets to actively anchor their specific evolutionary stages. Second, the **Temporal Transition Gating (TTG)** module enforces forward causality on the timeline. By employing asymmetric local temporal windows, TTG acts as a strict mathematical filter—mandating that an execution stage must be flanked by historical preparation and future completion. This effectively blocks static isolated noise and yields refined temporal action scores. Finally, the **Linguistic Knowledge Transfer (LKT)** module employs knowledge distillation and contrastive learning to embed this explicit structural knowledge from the refined scores into the student visual backbone.

In summary, the main contributions of this work are three-fold:

- We propose a novel perspective for WS-TAL that reconceptualizes actions as multi-stage action evolutions. By leveraging linguistic priors from LLMs, we successfully introduce

explicit structural and causal constraints to mitigate temporal ambiguity.

- We propose the LTKT framework, which consists of three interconnected modules (TSG, TTG, and LKT). Notably, the TTG module physically blocks temporal noise through asymmetric causal gating, providing highly reliable supervision.
- Extensive experiments on standard benchmarks, including THUMOS14 and ActivityNet, demonstrate that LTKT achieves state-of-the-art performance.

2 Related Work

2.1 Vision-Language Models

The evolution of Vision-Language Models (VLMs) like CLIP (Radford et al., 2021), BLIP (Li et al., 2022b), and SigLIP (Zhai et al., 2023) has transformed cross-modal learning through large-scale pre-training. To extend these capabilities to the temporal domain, a new lineage of video-specific VLMs has emerged. Approaches such as VideoCLIP (Xu et al., 2021), ActionCLIP (Wang et al., 2023a), and VideoCLIP-XL (Wang et al., 2024) integrate multi-frame sampling strategies with temporal attention mechanisms to capture complex video-text dynamics, proving highly effective for comprehensive video retrieval and recognition tasks.

Despite their success in video-level tasks, applying VLMs to fine-grained temporal localization is challenging. Treating a video as a holistic event leads to a lack of necessary temporal sensitivity, making it difficult to precisely delineate action boundaries in untrimmed videos.

2.2 Weakly-Supervised Action Localization

Weakly-Supervised Temporal Action Localization (WS-TAL) aims to localize action instances in untrimmed videos using only video-level category labels. The dominant approach formulates this as a Multiple Instance Learning (MIL) problem (Zhang et al., 2021; Ji et al., 2021; Gao et al., 2022; Li et al., 2022a; Shi et al., 2022), where videos are processed as bags of snippets.

Purely Visual WS-TAL. To mitigate action-background confusion in early MIL paradigms, methods introduced contrastive learning (Lee et al., 2020; Li et al., 2024b) and auxiliary actionness modeling (Ma et al., 2021; Li et al., 2023b). However, relying solely on visual cues, these models

fundamentally struggle to distinguish actual actions from visually correlated static backgrounds.

Language-Enhanced WS-TAL. To enrich semantic representations, recent works incorporate language modalities. Ju *et al.* (Ju *et al.*, 2023) used collaborative distillation, while PVLR (Lim *et al.*, 2024) and Zhang *et al.* (Zhang *et al.*, 2025a) integrated VLM or MLLM knowledge. Although validating the potential of cross-modal information, these methods rely on macroscopic text-video matching. By treating actions as monolithic events, they ignore chronological progression, causing boundary over-suppression during ambiguous transitions.

Our Distinction. In stark contrast to existing methodologies, our LTKT fundamentally reframes actions into multi-stage action evolutions. Instead of relying on coarse holistic alignments, we introduce fine-grained linguistic priors to explicitly model the preparation, execution, and completion phases of an action. By coupling these multi-stage priors with an asymmetric causal gating mechanism, LTKT effectively breaks the visual ambiguity and achieves precise boundary localization.

3 Methodology

3.1 Overview

In Weakly-Supervised Temporal Action Localization (WS-TAL), we are given an untrimmed video $V = \{v_t\}_{t=1}^T$ with T snippets, and its corresponding video-level categorical label $\mathbf{y} \in \{0, 1\}^C$, where C is the total number of action categories. Our goal is to predict a set of action instances $\{(t_{start}, t_{end}, c, q)\}$, denoting the start time, end time, category, and confidence score, respectively.

To alleviate the severe temporal ambiguity caused by the lack of snippet-level annotations, we propose the Linguistic Temporal Knowledge Transfer (LTKT) Network. As illustrated in Figure 2, LTKT comprises a teacher-student architecture driven by three cohesive modules. First, the **Temporal Stage Grounding (TSG)** module introduces text-driven structural priors to anchor unconstrained visual snippets into specific evolutionary stages. Second, the **Temporal Transition Gating (TTG)** module applies asymmetric temporal windows to enforce forward causality, generating refined temporal action scores. Finally, the **Linguistic Knowledge Transfer (LKT)** module seamlessly transfers this structural knowledge into the student backbone via knowledge distillation

and contrastive learning.

3.2 Video Backbone and MIL Baseline

Given an untrimmed video V , we first divide it into T non-overlapping snippets. Following standard WS-TAL protocols, we employ a pre-trained I3D (Carreira and Zisserman, 2017) model as our primary student network to extract snippet-level visual features. These features are mapped by a base temporal convolutional neural network (CNN) to generate the unified backbone embeddings $F_{vb} \in \mathbb{R}^{T \times d_{vb}}$, where d_{vb} is the feature dimension. To equip the visual backbone with basic category-discriminative capabilities, a standard Multiple Instance Learning (MIL) branch is attached. Specifically, F_{vb} is fed into a temporal CNN and a linear classifier to produce the snippet-level class activation sequence $S \in \mathbb{R}^{T \times C}$. A temporal aggregation function (*e.g.*, Top- k pooling) is then applied to S to derive the video-level categorical prediction $\hat{y} \in \mathbb{R}^C$. The baseline is supervised via the cross-entropy loss: $L_{mil} = -\sum_{c=1}^C y_c \log(\hat{y}_c)$.

While L_{mil} captures the general presence of actions, the pure video backbone lacks explicit temporal constraints, easily overfitting to salient snippets and plunging into temporal ambiguity. To resolve this, we introduce an auxiliary VLM-based teacher branch to provide structural guidance.

3.3 Temporal Stage Grounding

To overcome the temporal ambiguity of the video backbone, the TSG module functions as a cross-modal teacher, leveraging pre-trained Vision-Language Models (VLMs) to explicitly locate multi-stage action evolutions.

Linguistic Prior Generation. For each action category $c \in \{1, \dots, C\}$, we prompt an LLM to generate descriptions for three strict temporal stages: *Preparation (pre)*, *Execution (core)*, and *Completion (post)*. We map these textual descriptions into the joint vision-language space using a pre-trained VLM text encoder, obtaining the temporal evolution dictionary $\mathbf{F}_{txt} \in \mathbb{R}^{C \times 3 \times d}$, where d is the feature dimension.

Cross-Modal Stage Anchoring. Given the input video, we extract snippet-level visual features $\mathbf{F}_{ve} \in \mathbb{R}^{T \times d}$ using the VLM vision encoder. To establish the dense alignment between visual observations and temporal priors, we project $\mathbf{F}_{ve}$ as the queries (Q), and the flattened dictionary $\mathbf{F}_{txt}$

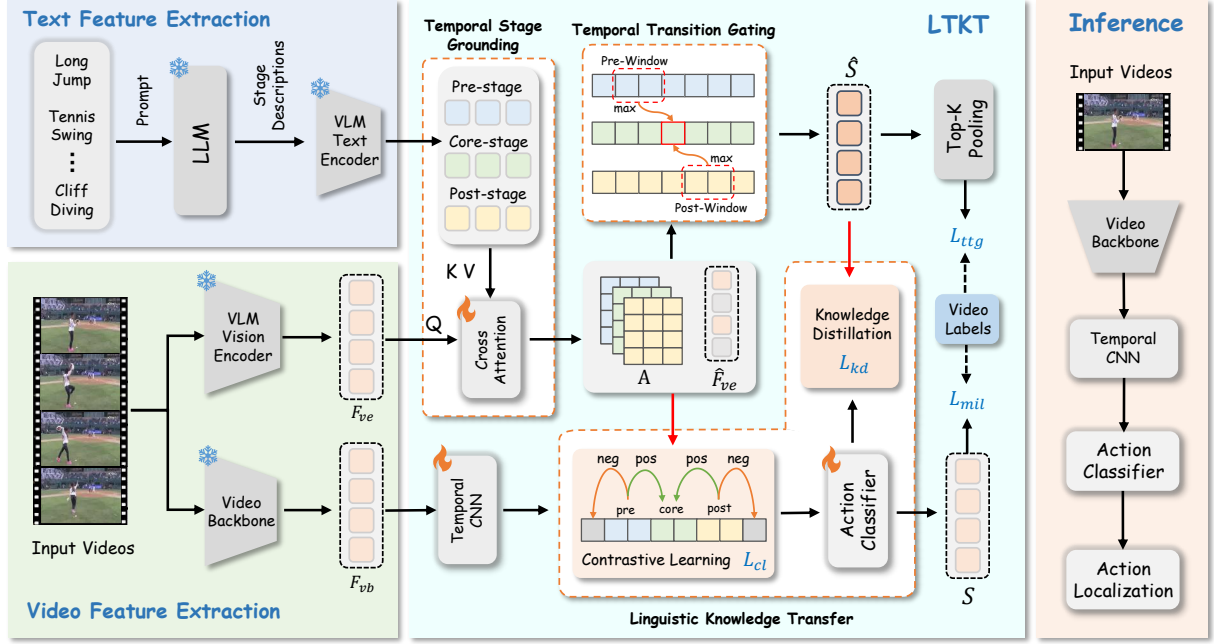


Figure 2: The overall architecture of the proposed LTKT framework. The LTKT leverages multi-stage linguistic priors to explicitly guide temporal localization through three core components: Temporal Stage Grounding, Temporal Transition Gating and Linguistic Knowledge Transfer.

as the keys (**K**) and values (**V**):

$$\mathbf{Q} = \mathbf{F}_{ve} \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}_{txt} \mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}_{txt} \mathbf{W}_V, \quad (1)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. We compute the cross-modal attention to yield the temporal semantic responses. By reshaping the normalized attention matrix, we obtain the fine-grained temporal stage attention $\mathbf{A} \in \mathbb{R}^{T \times C \times 3}$ and the fused features $\hat{\mathbf{F}}_{ve}$:

$$\mathbf{A} = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right), \quad \hat{\mathbf{F}}_{ve} = \mathbf{A}\mathbf{V} \quad (2)$$

Physically, each element $A_{t,c,s}$ represents the probability of the t -th visual snippet belonging to the s -th evolutionary stage ($s \in \{pre, core, post\}$) of action category c . Consequently, TSG effectively translates isolated visual pixels into structured temporal semantics, providing the fundamental stage awareness required for subsequent temporal gating.

3.4 Temporal Transition Gating

Although the TSG module endows visual features with temporal stage awareness, pure probability responses still struggle in complex weakly-supervised scenarios. To alleviate such temporal ambiguity, we introduce the Temporal Transition Gating (TTG) module. The core insight is that a genuine action must obey forward causality along

the timeline: a valid execution stage (*core*) must be preceded by a preparation stage (*pre*) and succeeded by a completion stage (*post*).

Asymmetric Directional Receptive Fields. Given the fine-grained stage attention tensor $\mathbf{A} \in \mathbb{R}^{T \times C \times 3}$ from the TSG module, we denote the probabilities for the three stages at snippet t for category c as $A_{t,c,pre}$, $A_{t,c,core}$, and $A_{t,c,post}$. To capture the directional transition contexts, we abandon symmetric global pooling—which blurs temporal boundaries—and instead construct directional temporal windows of size ω . We aggregate the historical (backward) and future (forward) contexts using 1D max-pooling for any stage $s \in \{pre, core, post\}$:

$$\begin{aligned} \overleftarrow{\Phi}_{t,c}^s &= \max_{j \in [t-\omega, t-1]} A_{j,c,s}, \\ \overrightarrow{\Phi}_{t,c}^s &= \max_{j \in [t+1, t+\omega]} A_{j,c,s}, \end{aligned} \quad (3)$$

where $\overleftarrow{\Phi}_{t,c}^s$ and $\overrightarrow{\Phi}_{t,c}^s$ represent the confidence of observing stage s in the immediate past and future neighborhoods, respectively.

Holistic Causal Gating and Refinement. We convert these aggregated directional contexts into non-linear gating signals via a sigmoid function $\sigma(\cdot)$. Based on our topological causality rules, we cross-modulate the raw probabilities to refine all three

stages simultaneously:

$$\hat{A}_{t,c,pre} = A_{t,c,pre} \cdot \sigma(\vec{\Phi}_{t,c}^{core}), \quad (4)$$

$$\hat{A}_{t,c,core} = A_{t,c,core} \cdot \sigma(\vec{\Phi}_{t,c}^{pre}) \cdot \sigma(\vec{\Phi}_{t,c}^{post}), \quad (5)$$

$$\hat{A}_{t,c,post} = A_{t,c,post} \cdot \sigma(\vec{\Phi}_{t,c}^{core}). \quad (6)$$

This acts as a strict set of logical “AND” gates along the temporal dimension. For instance, if an isolated high $A_{t,c,pre}$ activation lacks a succeeding execution phase ($\vec{\Phi}_{t,c}^{core} \rightarrow 0$), its gate closes, completely suppressing the static noise. Finally, the holistic temporal action score is seamlessly fused by aggregating the structurally validated stages:

$$\hat{S}_{t,c} = \hat{A}_{t,c,pre} + \hat{A}_{t,c,core} + \hat{A}_{t,c,post}, \quad (7)$$

where $\hat{S} \in \mathbb{R}^{T \times C}$ denotes the refined temporal action scores.

Teacher Branch Optimization. To ensure the VLM teacher branch correctly learns to ground the temporal stages, we aggregate the refined scores $\hat{S}$ using a Top- k pooling function to obtain the video-level prediction $\hat{y}^t \in \mathbb{R}^C$. The teacher branch is optimized via the cross-entropy loss: $L_{ttg} = -\sum_{c=1}^C y_c \log(\hat{y}_c^t)$.

3.5 Linguistic Knowledge Transfer

While the VLM teacher branch provides high-quality structural guidance, our ultimate goal is to empower the primary video backbone (student) to independently predict accurate action boundaries during inference. To seamlessly transfer this explicit structural knowledge, the LKT module employs two complementary mechanisms: decision-level distillation and feature-level contrastive regularization.

Linguistic Knowledge Distillation (LKD). Traditional distillation approaches often map low-confidence teacher scores into absolute hard negative labels (*i.e.*, zeros). However, in untrimmed videos, non-salient transition frames naturally exhibit high semantic ambiguity. Forcing strict hard negative labels inevitably leads to over-suppression and harms the boundary continuity.

To resolve this, we propose an asymmetric distillation strategy driven by a confidence threshold α . For a given snippet t , we generate the target probability distribution y_t^{ttg} based on the maximum response of the refined temporal scores $\hat{S}$:

$$y_{t,c}^{ttg} = \begin{cases} \mathbb{I}(\hat{S}_{t,c} > \alpha), & \text{if } \max_k(\hat{S}_{t,k}) > \alpha \\ \frac{1}{C}, & \text{if } \max_k(\hat{S}_{t,k}) \leq \alpha \end{cases} \quad (8)$$

where $\mathbb{I}(\cdot)$ is the indicator function. This asymmetric mapping behaves as a temporal uncertainty calibrator: when a snippet exhibits a strong temporal action score, we impart a sharp, low-entropy hard label to enforce strict boundary discrimination; conversely, when all class scores fall below α , the snippet is treated as an ambiguous background context. Instead of enforcing a rigid zero, we assign a uniform label distribution ($1/C$), which acts as a maximum entropy regularizer. This prevents the student network from becoming overconfident on complex backgrounds and preserves the structural smoothness of action evolutions. The student’s predictions S are then supervised by y^{ttg} via the cross-entropy loss, denoted as L_{lkd} .

Temporal Contrastive Learning (TCL). While LKD focuses on snippet-level classification, we further introduce TCL to regularize the student’s representation space $\mathbf{F}_{vb}$. Specifically, we explicitly encode the temporal causality of action evolution into the feature manifold using directional contrastive learning.

Instead of directly reusing the raw attention, we calculate the fine-grained semantic similarity between the fused features $\hat{\mathbf{F}}_{ve}$ and the temporal textual dictionary $\mathbf{F}_{txt}$. We denote this normalized similarity tensor as $\mathbf{U} \in \mathbb{R}^{T \times C \times 3}$, where each element $U_{t,c,s}$ represents the semantic alignment score for the s -th evolutionary stage ($s \in \{pre, core, post\}$) of class c at snippet t . The background probability is subsequently derived as $P_{bg}^{(t)} = 1 - \sum_s U_{t,c,s}$. For a given snippet i , the similarity with snippet j in the student feature space is defined as $S_{i,j} = \exp(\mathbf{F}_{vb}^{(i)} \cdot \mathbf{F}_{vb}^{(j)} / \tau)$, where τ is the temperature. We define directional temporal masks: $\mathcal{M}_f(i, j) = \mathbb{I}(0 < j - i \leq \omega)$ and $\mathcal{M}_p(i, j) = \mathbb{I}(-\omega \leq j - i < 0)$, where ω is the window size.

Taking the *pre* stage as an anchor, the causal logic dictates that it should be temporally close to the *core* execution in the future (positive samples) and far from the *bg* context in the past (negative samples). The directional contrastive loss for the *pre* stage is formulated as:

$$L_{pre} = -\frac{1}{T} \sum_{i=1}^T U_i \times \log \frac{\sum_j S_{i,j} U_j \mathcal{M}_f}{\sum_j S_{i,j} (U_j \mathcal{M}_f + P_{bg}^{(j)} \mathcal{M}_p)}. \quad (9)$$

Similarly, the *post* stage seeks *core* features in

the past and pushes away bg features in the future, yielding L_{post} . The total temporal contrastive loss is $L_{tcl} = (L_{pre} + L_{post})/2$. This directional regularization ensures that the learned visual features encapsulate the evolutionary progression.

3.6 Overall Objective

Our LTKT is trained in an end-to-end manner. The overall multi-task learning objective is defined as:

$$L_{total} = L_{mil} + \lambda_1 L_{ttg} + \lambda_2 L_{tcl} + \lambda_3 L_{lkd}, \quad (10)$$

where λ_1 , λ_2 and λ_3 are balancing hyperparameters.

During inference, the heavy teacher framework (TSG, TTG, and LKT) is entirely discarded, and only the lightweight video backbone is utilized to generate the final temporal action localization proposals, ensuring high computational efficiency.

4 Experiment

In this section, we conduct extensive experiments. For more experimental results, such as hyperparameter sensitivity, noise robustness, efficiency evaluation, *etc.*, please refer to the **Appendix**.

Table 1: Performance comparison with SOTA methods on the THUMOS14 dataset. Rows highlighted in green denote VLM-driven methods.

Method	mAP@IoU (%)				Avg. 0.1:0.7
	0.1	0.3	0.5	0.7	
AICL (Li et al., 2023b)	73.1	58.2	36.9	14.9	46.4
Ren et al. (Ren et al., 2023)	71.8	58.9	40.0	15.1	47.0
Zhou et al. (Zhou et al., 2023)	74.0	60.7	42.7	13.1	48.3
PivoTAL (Rizve et al., 2023)	74.1	61.7	42.8	16.7	49.6
ISSF (Yun et al., 2024)	72.4	58.4	41.8	12.8	46.8
Hu et al. (Hu et al., 2024)	74.1	60.0	41.1	15.1	48.2
SDANet (Zhou et al., 2025)	72.9	58.6	40.6	15.8	47.6
NoCo (Zhang et al., 2025b)	75.2	63.0	44.0	17.7	50.9
Liu et al. (Liu and Liu, 2025)	76.8	65.4	44.1	17.1	51.9
Li et al. (Li et al., 2023a)	-	56.2	39.3	15.2	-
Ju et al. (Ju et al., 2023)	73.5	61.5	42.0	16.8	49.4
PVLR (Lim et al., 2024)	74.9	61.4	45.1	17.1	50.3
Zhang et al. (Zhang et al., 2025a)	74.3	61.8	43.0	16.6	49.8
MSLU (Shao et al., 2026)	76.1	62.9	45.8	17.7	51.5
LTKT (Ours)	78.0	68.4	46.1	19.8	54.2

4.1 Datasets and Evaluation

THUMOS14 (Idrees et al., 2017) contains 200 videos in the validation set and 213 videos in the test set with 20 action categories. ActivityNet1.3 (Caba Heilbron et al., 2015) contains 10,024 training videos, 4,926 untrimmed validation videos, and 5,044 test videos, for which the action-instance annotations are withheld. The dataset contains 200 action categories.

We use the mean Average Precision (mAP) with different temporal Intersection over Union (t-IoU) thresholds to evaluate the performance. The t-IoU thresholds for THUMOS14 are [0.1:0.1:0.7], and [0.5:0.05:0.95] for ActivityNet.

4.2 Implementation Details

We employ I3D (Carreira and Zisserman, 2017) as the video backbone, leverage VideoCLIP-XL (Wang et al., 2024) as the pre-trained VLM, and adopt Qwen-3.5 (Qwen Team, 2026) as the default LLM to generate multi-stage linguistic descriptions. We optimize the network using learning rates of $1e-4$ and $1e-5$ for THUMOS14 and ActivityNet, respectively. Consistent with the baseline, we set $T = 750$, $k = T/8$ for THUMOS14 and $T = 50$, $k = T/4$ for ActivityNet1.2 and 1.3. The new hyperparameters introduced by the LTKT remain consistent across all datasets, $\lambda_1 = 0.1$, $\lambda_2 = 0.6$, $\lambda_3 = 0.5$, $\alpha = 0.9$. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.3 Comparison with SOTA Methods

Performance on THUMOS14. Table 1 shows the performance on THUMOS14. Specifically, LTKT yields an impressive average mAP (0.1:0.7) of 54.2%, outperforming the most recent methods such as Liu et al. (Liu and Liu, 2025) (51.9%) and NoCo (Zhang et al., 2025b) (50.9%). Notably, at the highly strict metric of mAP@0.7, our method achieves 19.8%, heavily surpassing previous paradigms that typically struggle with precise boundary localization. Furthermore, when compared to contemporary VLM-driven methods (*e.g.*, PVLR (Lim et al., 2024) and Zhang et al. (Zhang et al., 2025a)) that typically treat actions as holistic events, LTKT achieves an absolute improvement of +3.9% in average mAP. This improvement supports our core hypothesis: explicitly modeling multi-stage temporal evolutions and enforcing causal gating provides far more actionable structural priors than cross-modal matching.

Performance on ActivityNet1.3. As reported in Table 2, LTKT achieves a new record on ActivityNet1.3. It yields an average mAP of 32.8%, outperforming the latest approach (Liu and Liu, 2025) by an absolute margin of 3.8%. Consistent improvements across both loose (mAP@0.5: 50.7%) and strict (mAP@0.95: 7.8%) thresholds demonstrate its superior boundary localization capability.

Table 2: Performance comparison with SOTA methods on the ActivityNet1.3 dataset. Rows highlighted in green denote VLM-driven methods.

Method	mAP@IoU (%)			
	0.5	0.75	0.95	Avg.
Huang <i>et al.</i> (Huang <i>et al.</i> , 2022)	40.6	24.6	5.9	25.0
ASM-Loc (He <i>et al.</i> , 2022)	41.0	24.9	6.2	25.1
DGCNN (Shi <i>et al.</i> , 2022)	37.2	23.8	5.8	23.9
Ren <i>et al.</i> (Ren <i>et al.</i> , 2023)	41.8	25.4	5.2	25.5
Wang <i>et al.</i> (Wang <i>et al.</i> , 2023b)	41.8	25.7	6.5	26.3
PivoTAL (Rizve <i>et al.</i> , 2023)	45.1	28.2	5.0	28.1
ISSF (Yun <i>et al.</i> , 2024)	39.4	25.8	6.4	25.8
Liu <i>et al.</i> (Liu and Liu, 2025)	46.9	28.6	6.5	29.0
Li <i>et al.</i> (Li <i>et al.</i> , 2023a)	41.8	26.0	6.0	26.0
PVLR (Lim <i>et al.</i> , 2024)	43.6	27.4	6.5	27.4
MSLU (Shao <i>et al.</i> , 2026)	43.8	27.8	6.6	27.5
LTKT (Ours)	50.7	32.8	7.8	32.8

Table 3: Ablation study of each component in LTKT. Experiments are conducted on THUMOS14.

TSG	TTG	LKT		mAP@IoU(%)				
		LKD	TCL	0.1	0.3	0.5	0.7	Avg.
				71.4	57.1	35.9	13.7	45.1
✓		✓		73.2	61.7	41.7	17.3	49.2
✓			✓	73.6	61.3	40.7	16.6	48.8
✓	✓	✓		77.1	67.1	45.4	19.8	53.5
✓	✓		✓	75.1	63.3	41.0	16.5	49.8
✓		✓	✓	74.0	62.2	41.7	16.9	49.6
✓	✓	✓	✓	78.0	68.4	46.1	19.8	54.2

4.4 Ablation Study

Effectiveness of Key Components. Table 3 details the performance trajectory when incrementally integrating our core modules into the pure-vision MIL baseline. The baseline struggles with severe temporal ambiguity, yielding only 45.1% in average mAP. Integrating the TSG to guide the LKD or TCL directly brings noticeable gains, verifying that linguistic stage priors provide valuable structural cues. However, pure stage probabilities still contain background noise. Introducing the TTG alongside TSG and LKD leads to a significant increase to 53.5% in average mAP. This demonstrates that filtering out static background noise through explicit causal gating is crucial before transferring knowledge to the student. Finally, by combining all components, our full LTKT framework achieves the optimal average mAP of 54.2%, proving the synergistic effect of decision-level distillation and feature-level structural regularization.

Analysis on Stage Combinations To examine the applicability of different temporal stage structures across action categories, we conduct a plug-and-play ablation over different stage combinations and

Table 4: Average Precision at IoU=0.5 for different action types under various stage combinations on THU-MOS14.

Stage	LongJump	CleanAndJerk	CricketShot	FrisbeeCatch
Core	81.6	40.7	16.3	16.4
Pre + Core	86.5	42.9	15.4	19.7
Core + Post	85.9	39.0	15.5	15.2
Full stage	88.3	46.9	15.2	16.1

Table 5: Ablation of the temporal gating strategy in the TTG module on THUMOS14. All models are evaluated under the same teacher-student framework.

Gating Strategy	mAP @ IoU (%)				
	0.1	0.3	0.5	0.7	Avg.
w/o Gating ($A_{t,c,core}$)	77.1	66.3	43.7	18.6	52.5
Avg Pooling ($A_{t,c,:}$)	75.8	62.9	39.7	16.5	49.5
Max Pooling ($A_{t,c,:}$)	77.4	66.1	42.9	17.4	51.9
TTG (Causal Gating)	78.0	68.4	46.1	19.8	54.2

report the per-class AP at IoU= 0.5 in Table 4. The results reveal clear action-type-dependent differences. For instantaneous actions, such as *CricketShot* and *FrisbeeCatch*, a three-stage decomposition can be redundant and may introduce adjacent background noise; consequently, configurations with fewer stages achieve better performance. In contrast, for actions involving multi-stage processes, such as *LongJump* and *CleanAndJerk*, incorporating more stages is beneficial. In particular, the full “Pre + Core + Post” structure imposes stronger causal constraints and achieves the best localization performance.

Analysis of Temporal Transition Gating. In Table 5, we investigate the necessity of the asymmetric causal gating within the TTG module. If we entirely remove the gating mechanism (“w/o Gating”) and solely rely on the raw execution (*core*) stage probability, the model achieves 52.5% average mAP. Applying traditional symmetric operations, such as Avg Pooling or Max Pooling across all stage probabilities, fails to yield improvements and even deteriorates performance (dropping to 49.5% with Avg Pooling). This degradation occurs because symmetric pooling is structure-agnostic; it smears stage boundaries and completely ignores the chronological progression of an action. In contrast, our proposed Causal Gating explicitly enforces the forward “Preparation → Execution → Completion” logic, acting as a strict mathematical filter to eliminate isolated false positives.

Exploration of Distillation Strategies. Table 6

Table 6: Ablation study on distillation strategies in the LKD module.

Distillation Strategy	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	Avg.
MSE	76.7	66.0	43.2	18.2	52.0
MAE	77.1	65.6	43.0	18.5	51.9
KL-Divergence	77.6	68.1	45.5	19.4	53.7
Ours	78.0	68.4	46.1	19.8	54.2

Table 7: Robustness analysis across different LLMs.

LLM	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	Avg.
Gemini-3	77.7	67.9	45.3	19.6	53.8
GPT-5.3	78.2	67.9	45.7	20.0	54.1
Grok-4	77.6	67.5	45.6	19.4	53.6
Kimi-2.5	77.7	67.5	45.0	19.1	53.4
DeepSeek-V3.2	77.4	67.4	45.9	19.6	53.6
Qwen-3.5	78.0	68.4	46.1	19.8	54.2

compares our proposed distillation strategy in the LKD module against standard knowledge transfer loss functions, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and Kullback-Leibler (KL) Divergence. Direct numerical regression losses like MSE (52.0%) and MAE (51.9%) treat the teacher’s scores as absolute continuous values, making the student overly sensitive to the teacher’s amplitude fluctuations and calibration errors. KL-Divergence serves as a strong baseline (53.7%) by matching the continuous probability distributions. However, our strategy (54.2%) outperforms all alternatives. By mapping confident predictions into sharp one-hot labels while employing a uniform distribution (maximum entropy) for ambiguous background snippets, our formulation successfully enforces strict boundary discrimination without suffering from the severe over-suppression typical of hard-zero assignments.

Robustness Across Different LLMs. A common concern in language-driven frameworks is the over-reliance on a specific proprietary language model. To demonstrate that the success of LTKT is not merely a byproduct of a specific LLM’s linguistic phrasing, we evaluate the framework using various LLMs (Qwen Team, 2026; DeepSeek-AI, 2025; Team et al., 2026, 2023; Achiam et al., 2023). As shown in Table 7, the performance remains remarkably stable across all tested models, encompassing both open-weight architectures and proprietary models. The average mAP exhibits minimal fluctu-

ation within an extremely tight window from 53.4% to 54.2%. This stability suggests that the method’s performance derives from the architectural design of temporal alignment and causal gating, rather than the specific prompting characteristics of an individual language model.

4.5 Qualitative Results

Figure 3 visualizes the temporal activation profiles of our LTKT against the baseline model. For complex actions like *CleanAndJerk*, the baseline suffers from severe over-suppression, ignoring the preparatory squatting phase and activating only during the highly discriminative climax (overhead lift). By explicitly modeling the multi-stage temporal evolution, LTKT seamlessly captures the complete action progression. Furthermore, in context-heavy instances like *BaseballPitch*, the baseline is easily distracted by visually similar surrounding frames, producing low overall scores and multiple false-positive spikes. In contrast, LTKT produces distinct boundaries with low background activation. This qualitative comparison indicates that our causal gating mechanism helps filter out contextual interference while maintaining structural completeness.

False Positive Analysis. To further investigate the source of our performance gains, we conduct a detailed error diagnosis using the DETAD (Alwassel et al., 2018) tool on THUMOS14, as shown in Figure 4. LTKT achieves a remarkable True Positive rate of 53.8%, outperforming the baseline by an absolute margin of 7.4%. This substantial boost is primarily driven by the drastic reduction in two major error types: Background Error drops significantly from 17.4% to 11.8%, and Localization Error decreases from 31.7% to 28.9%. These metrics quantitatively confirm that our proposed causal gating explicitly filters out static background noise, while the temporal contrastive learning successfully mitigates temporal ambiguity to produce precise action boundaries.

5 Conclusion

In this paper, we present LTKT, which fundamentally reframes actions into multi-stage action evolutions to break severe temporal ambiguity. Specifically, TSG extracts fine-grained linguistic stage priors, while TTG explicitly filters static backgrounds via causal gating to prevent boundary over-suppression. To ensure efficient inference, LKT transfers this structural knowledge to a lightweight

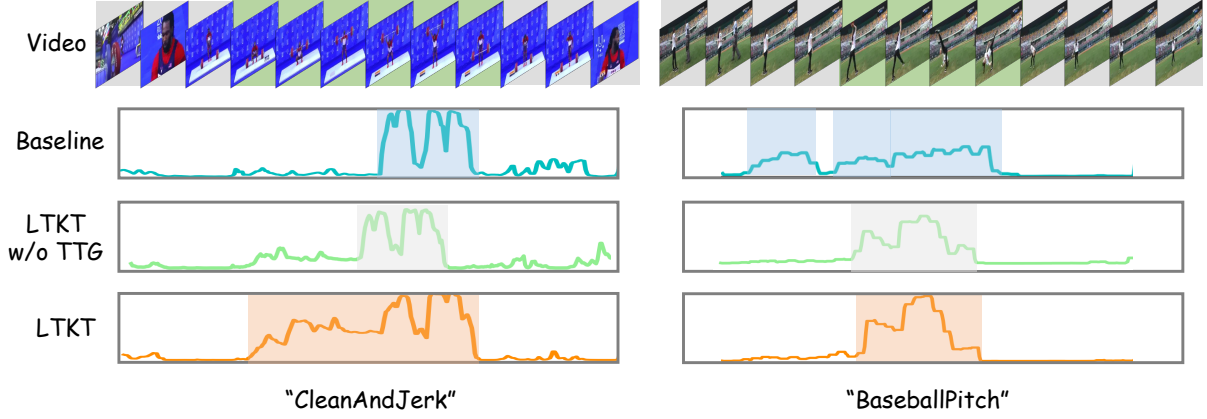


Figure 3: Qualitative results. The green background areas in the video represent the ground truth.

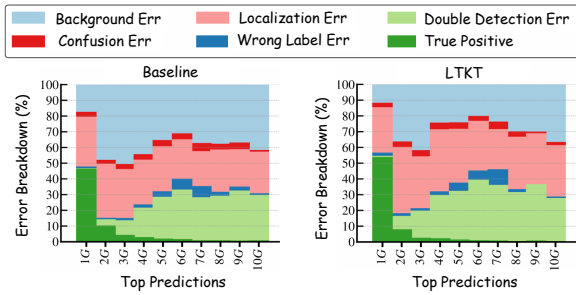


Figure 4: False positive analysis using DETAD. G is the number of ground truth instances.

student backbone through distillation and contrastive learning. Extensive experiments demonstrate that LTKT achieves SOTA performance, offering a new baseline for VLM-driven Weakly-Supervised Temporal Action Localization.

Limitations and Future Work

While the proposed Linguistic Temporal Knowledge Transfer (LTKT) framework demonstrates competitive performance in weakly-supervised temporal action localization, we acknowledge several limitations that provide avenues for future research:

Rigid Three-Stage Transition Assumption. Our Temporal Stage Grounding (TSG) module explicitly prompts the LLM to decompose each action into exactly three continuous and sequential stages: Preparation, Execution, and Completion. We adopt this fixed tripartite design to provide a simple and intuitive validation of our core idea, namely, the incorporation of structured linguistic priors. Although this structure effectively captures the temporal characteristics of many sports actions, such as *LongJump*, it may be insufficient for more com-

plex human activities. For example, daily activities such as cooking or applying makeup do not always conform naturally to a rigid three stage temporal progression. Moreover, our per class analysis indicates that instantaneous actions may benefit from fewer stages.

A promising direction for future work is to enable the LLM to adaptively generate action specific stage structures by determining the appropriate number of sub stages for each action category. This extension could improve the flexibility of the framework and facilitate its generalization to a broader range of diverse daily activities.

Fixed Temporal Window for Causal Gating. To enforce forward causality, the Temporal Transition Gating (TTG) module relies on a predefined local temporal window size, denoted as ω , to aggregate historical and future contexts. However, in real world untrimmed videos, action durations can vary substantially across categories and even among instances within the same category. As a result, a fixed window size ω may fail to capture the full causal dependencies of particularly long actions, while introducing excessive background noise for very short actions.

Although the fixed window strategy remains effective under common temporal variations, it may be less suitable for actions with highly diverse temporal scales. Therefore, developing a dynamic temporal window mechanism that can adapt to action duration is a promising direction for future work. Such a mechanism is particularly challenging in weakly supervised settings, where fine grained temporal boundary annotations are unavailable.

6 Acknowledgements

This work is supported by the Anhui Province Key Research and Development Plan (No. 202304a05020045), National Natural Science Foundation of China (No. 62676382, No. 62406098 and No. 62376256) and the Joint Fund for Medical Artificial Intelligence (No. MAI2022Q011).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. 2018. Diagnosing error in temporal action detectors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 256–272.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970. IEEE.
- Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.
- Mengyuan Chen, Junyu Gao, and Changsheng Xu. 2023. Uncertainty-aware dual-evidential learning for weakly-supervised temporal action localization. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15896–15911.
- Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. 2022. Dual-evidential learning for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 192–208. Springer.
- DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.
- Junyu Gao, Mengyuan Chen, and Changsheng Xu. 2022. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009.
- 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*, pages 13925–13935.
- 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*, pages 1591–1599.
- Xin Hu, Kai Li, Deep Patel, Erik Kruus, Martin Renqiang Min, and Zhengming Ding. 2024. Weakly-supervised temporal action localization with multi-modal plateau transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2713.
- 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*, pages 3272–3281.
- Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorbunov, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos “in the wild”. *CVIU*, pages 1–23.
- Yuan Ji, Xu Jia, Huchuan Lu, and Xiang Ruan. 2021. Weakly-supervised temporal action localization via cross-stream collaborative learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 853–861.
- 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*, pages 14751–14762.
- Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. 2020. Background suppression network for weakly-supervised temporal action localization. In *AAAI*.
- Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. 2023a. [Boosting weakly-supervised temporal action localization with text information](#). *Preprint*, arXiv:2305.00607.
- Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. 2022a. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19914–19924.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Zhilin Li, Zilei Wang, and Cerui Dong. 2024a. Multi-level semantic and adaptive actionness learning for weakly supervised temporal action localization. *Neural Networks*, page 106905.

- Zhilin Li, Zilei Wang, and Qinying Liu. 2023b. Actionness inconsistency-guided contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1513–1521.
- Zhilin Li, Zilei Wang, and Qinying Liu. 2024b. Weakly supervised temporal action localization with actionness-guided false positive suppression. *Neural Networks*, 175:106307.
- Ziqiang Li, Yongxin Ge, Jiaruo Yu, and Zhongming Chen. 2022c. Forcing the whole video as background: An adversarial learning strategy for weakly temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5371–5379.
- 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*, pages 5507–5516.
- Ziyi Liu and Yangcen Liu. 2025. Bridge the gap: From weak to full supervision for temporal action localization with pseudoformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8711–8720.
- Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. 2021. Weakly supervised action selection learning in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7596.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- 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*, pages 2394–2404.
- 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*, pages 22992–23002.
- Yuxiang Shao, Feifei Zhang, and Changsheng Xu. 2026. [Modeling semantic and localization uncertainty for weakly supervised temporal action localization](#). *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.
- Haichao Shi, Xiao-Yu Zhang, Changsheng Li, Lixing Gong, Yong Li, and Yongjun Bao. 2022. Dynamic graph modeling for weakly-supervised temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3820–3828.
- 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 (ICCV)*, pages 6622–6632.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. [Kimi k2.5: Visual agentic intelligence](#). *Preprint*, arXiv:2602.02276.
- Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. 2024. Videoclip-xl: Advancing long description understanding for video clip models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16061–16075.
- Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. 2023a. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE transactions on neural networks and learning systems*, 36(1):625–637.
- Yu Wang, Yadong Li, and Hongbin Wang. 2023b. 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*, pages 18878–18887.
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800.
- 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*, volume 38, pages 6908–6916.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

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*, pages 16010–16019.

Quan Zhang, Jinwei Fang, Rui Yuan, Xi Tang, Yuxin Qi, Ke Zhang, and Chun Yuan. 2025a. Weakly supervised temporal action localization via dual-prior collaborative learning guided by multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24139–24148.

Quan Zhang, Yuxin Qi, Xi Tang, Rui Yuan, Xi Lin, Ke Zhang, and Chun Yuan. 2025b. Rethinking pseudo-label guided learning for weakly supervised temporal action localization from the perspective of noise correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10085–10093.

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*, pages 23003–23012.

Wei Zhou, Kang Lin, Weipeng Hu, Chao Xie, Tao Su, Haifeng Hu, and Yap-Peng Tan. 2025. Snippet-inter difference attention network for weakly-supervised temporal action localization. *IEEE Transactions on Multimedia*.

A Appendix

A.1 Performance on ActivityNet1.2

Table 8 details the comparative results on the ActivityNet1.2 dataset, which features diverse and complex untrimmed videos with highly variable action durations. Our LTKT consistently establishes new records among all weakly-supervised approaches. LTKT achieves an average mAP of 34.6%, significantly outperforming the latest pure-vision method NoCo (Zhang et al., 2025b) (30.7%) and the VLM-based paradigm Zhang et al. (Zhang et al., 2025a) (30.1%) by absolute margins of +3.9% and +4.5%, respectively. The massive improvements on these challenging benchmarks strongly demonstrate the

Table 8: Performance comparison with SOTA methods on the ActivityNet1.2 dataset. Rows highlighted in green denote VLM-driven methods.

Method	mAP@IoU (%)			
	0.5	0.75	0.95	Avg.
CoLA (Zhang et al., 2021)	42.7	25.7	5.8	26.1
CO2-Net (Hong et al., 2021)	43.3	26.3	5.2	26.4
Li et al. (Li et al., 2022c)	41.6	24.8	5.4	25.2
DELU (Chen et al., 2022)	44.2	26.7	5.4	26.9
DDG-Net (Tang et al., 2023)	44.3	26.9	5.5	27.0
Ren et al. (Ren et al., 2023)	44.2	26.1	5.3	26.5
Hu et al. (Hu et al., 2024)	45.1	27.7	5.5	27.6
NoCo (Zhang et al., 2025b)	49.1	30.5	6.7	30.7
Ju et al. (Ju et al., 2023)	48.3	29.3	6.1	29.6
Zhang et al. (Zhang et al., 2025a)	48.3	30.1	6.8	30.1
LTKT (Ours)	55.7	34.4	7.7	34.6

exceptional robustness and generalizability of transferring multi-stage temporal priors to resolve temporal ambiguity.

A.2 Hyperparameter Robustness Evaluation

Table 9: Sensitivity of λ_1 . The AVG means average mAP under multiple t-IoU thresholds as {0.1:0.1:0.7} on THUMOS14.

λ_1	mAP@t-IoU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
0.05	78.2	74.3	68.3	59.2	45.5	32.1	20.1	54.0
0.1	78.0	74.5	68.4	59.4	46.1	33.3	19.8	54.2
0.2	77.7	74.4	68.0	58.8	45.5	31.2	19.2	53.5
0.4	76.8	73.4	67.2	57.6	44.8	31.9	19.1	53.0
0.8	77.8	74.0	67.6	57.2	44.3	30.7	18.5	52.9

Table 10: Sensitivity of λ_2 . The AVG means average mAP under multiple t-IoU thresholds as {0.1:0.1:0.7} on THUMOS14.

λ_2	mAP@t-IoU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
0.1	77.8	74.6	67.8	59.3	45.4	32.5	19.3	53.8
0.3	78.0	74.4	67.9	59.3	45.5	32.5	19.3	53.8
0.6	78.0	74.5	68.4	59.4	46.1	33.3	19.8	54.2
0.8	78.0	74.4	68.0	59.2	45.4	32.9	19.2	53.9
1.0	77.6	74.1	67.2	59.0	45.6	32.4	19.7	53.7

Impact of Loss Weights (λ_1 , λ_2 , and λ_3). We first evaluate the robustness of our LTKT framework to the loss balancing weights, as summarized in Tables 9, 10, and 11. The experiments are conducted on the THUMOS14 dataset. Overall, the model demonstrates remarkable stability across a wide range of weight configurations. For λ_1 , the performance peaks at 0.1 with an average mAP of

Table 11: Sensitivity of λ_3 . The AVG means average mAP under multiple t-IOU thresholds as $\{0.1:0.1:0.7\}$ on THUMOS14.

λ_3	mAP@t-IOU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
0.3	77.7	74.4	67.9	59.1	45.9	31.9	19.1	53.7
0.5	78.0	74.5	68.4	59.4	46.1	33.3	19.8	54.2
0.7	77.7	74.3	67.6	59.2	45.2	31.6	19.2	53.6
1.0	77.3	74.3	67.7	59.8	45.1	31.6	19.6	53.6
1.2	77.2	74.0	67.4	59.1	44.6	31.9	19.5	53.4

Table 12: Sensitivity of confidence threshold α . The AVG means average mAP under multiple t-IOU thresholds as $\{0.1:0.1:0.7\}$ on THUMOS14.

α	mAP@t-IOU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
0.85	79.1	74.6	67.0	56.6	43.2	30.8	19.0	52.9
0.88	78.1	73.9	67.3	58.2	44.7	32.2	19.6	53.4
0.90	78.0	74.5	68.4	59.4	46.1	33.3	19.8	54.2
0.92	77.2	73.9	67.1	58.5	45.5	32.0	19.1	53.3
0.95	76.8	72.8	66.1	57.2	44.3	31.6	18.3	52.5

54.2%. While excessively large values (*e.g.*, 0.8) cause a slight degradation to 52.9%, the framework remains highly competitive. Similarly, λ_2 and λ_3 exhibit minimal sensitivity, maintaining an average mAP comfortably above 53.4% across almost all tested settings. The optimal combination is empirically found at $\lambda_1 = 0.1$, $\lambda_2 = 0.6$, and $\lambda_3 = 0.5$. This widespread robustness indicates that the performance gains of LTKT are fundamentally derived from the architectural design of the temporal gating and knowledge transfer, rather than heavy reliance on exhaustive hyperparameter tuning.

Sensitivity of Confidence Threshold α . Table 12 investigates the impact of the confidence threshold α , which controls the strictness of assigning pseudo-labels during the Linguistic Knowledge Transfer (LKT) process. We observe a clear trade-off between proposal recall and localization precision. When α is set too low (*e.g.*, 0.85), the model achieves a high mAP@0.1 of 79.1% but suffers a severe drop at stricter metrics (*e.g.*, dropping to 43.2% at mAP@0.5). This occurs because a loose threshold erroneously converts ambiguous transition frames into hard labels, reintroducing temporal noise. Conversely, an overly strict threshold (*e.g.*, $\alpha = 0.95$) filters out too many valid temporal cues, causing the student model to underfit and dropping the average mAP to 52.5%. The threshold $\alpha = 0.90$ strikes the optimal balance, successfully

Table 13: Sensitivity of the window size ω . The AVG means average mAP under multiple t-IOU thresholds as $\{0.1:0.1:0.7\}$ on THUMOS14.

ω	mAP@t-IOU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
5	77.8	74.4	68.2	59.5	45.6	32.4	19.2	53.9
10	78.0	74.5	67.9	59.4	45.0	31.9	19.4	53.7
15	78.0	74.5	68.4	59.4	46.1	33.3	19.8	54.2
20	77.9	74.4	67.8	59.1	45.2	32.2	19.4	53.7
30	77.7	74.3	68.0	59.3	45.6	32.6	19.3	53.8

filtering out contextual ambiguity while preserving sufficient structural guidance to achieve the peak average mAP of 54.2%.

Sensitivity of the window size ω . In Table 13, we thoroughly evaluate the impact of the local temporal window size ω , which determines the receptive field for aggregating historical (pre-stage) and future (post-stage) transitional actions. Overall, the model demonstrates remarkable robustness across different settings of ω . The average mAP (AVG) on the THUMOS14 dataset fluctuates within a tight margin of 53.7% to 54.2%, indicating that our proposed framework does not rely heavily on exhaustive hyperparameter tuning. Specifically, the model achieves its optimal performance at $\omega = 15$, yielding an AVG mAP of 54.2% and peaking at strict t-IOU thresholds (*e.g.*, 46.1% at mAP@0.5 and 19.8% at mAP@0.7). When the window is too small (*e.g.*, $\omega = 5$ or 10), the restricted receptive field struggles to encapsulate the complete causal dependencies of prolonged actions, leading to a slight degradation in capturing the full temporal evolutions. Conversely, an excessively large window (*e.g.*, $\omega = 20$ or 30) inevitably incorporates irrelevant static background noise or mistakenly aggregates fragments from adjacent action instances, thereby blurring the precise temporal boundaries. Therefore, $\omega = 15$ strikes the optimal balance between providing sufficient causal context and maintaining sharp boundary discrimination.

A.3 Noise Robustness Evaluation

To comprehensively evaluate the stability and generalizability of our proposed LTKT framework against input perturbations, we conduct a rigorous noise robustness test. Specifically, we inject zero-mean Gaussian noise with varying standard deviations directly into the extracted visual features during the inference phase on the THUMOS14 dataset.

Table 14: Robustness evaluation of LTKT under varying levels of visual feature perturbation. The AVG means average mAP under multiple t-IoU thresholds as {0.1:0.1:0.7} on THUMOS14.

std	mAP@t-IoU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
0	78.0	74.5	68.4	59.4	46.1	33.3	19.8	54.2
0.1	78.0	74.5	68.3	59.3	45.8	32.9	19.6	54.0
0.2	78.0	74.3	68.2	59.2	45.7	32.7	19.5	53.9
0.3	77.9	74.2	68.1	58.7	45.7	32.7	19.3	53.8
0.4	77.7	74.0	67.8	58.3	45.3	32.1	19.0	53.4

As demonstrated in Table 14, LTKT is robust to visual feature perturbations. Even when the standard deviation of the injected noise increases to 0.4, the average mAP experiences only a minor degradation, stabilizing consistently between 53.4% and 54.0%. Even under the most severe noise condition ($std = 0.4$), the overall performance drop is a mere 0.8% in average mAP compared to the clean baseline (54.2%). This remarkable robustness provides strong empirical validation for our core architectural hypothesis.

A.4 Efficiency Evaluation

During inference, the heavy teacher framework (TSG, TTG, and LKT) is entirely omitted. The final temporal localization proposals are generated solely by the lightweight video backbone, facilitating highly efficient real-world deployment. Table 15 provides a comprehensive comparison of computational complexity (GFLOPs, Parameters), inference latency, and detection performance (AVG mAP) on the THUMOS14 dataset. All experiments are conducted on a single RTX 3090 GPU. Compared to the purely visual baseline, our LTKT framework achieves a remarkable absolute gain of +9.1% in average mAP, while incurring only a marginal increase in computational overhead (+2.3 GFLOPs, +4.1M parameters, and a negligible +0.6ms in inference time).

More importantly, when compared to recent state-of-the-art methods, LTKT demonstrates an exceptional performance-efficiency trade-off. For instance, while MSLU achieves a competitive 51.5% mAP, it requires heavily parameterized modules (28.1M parameters) and massive computational cost (231 GFLOPs). In stark contrast, LTKT reaches a much higher mAP of 54.2% with only a fraction of the computational burden (7.0 GFLOPs and 10.4M parameters) and runs at highly efficient

speeds (47.7ms). This extreme efficiency is directly attributed to our Linguistic Knowledge Transfer (LKT) paradigm. By distilling the heavy multi-stage linguistic priors and causal gating logic into a lightweight student backbone during training, LTKT completely circumvents the need for any cumbersome LLM or VLM components during the inference phase, making it highly robust for real-world, resource-constrained deployments.

A.5 Comprehensive False Positive Analysis

To provide a deeper insight into the source of our performance improvements, we conduct an exhaustive error diagnosis using the DETAD (Alwassel et al., 2018) tool on the THUMOS14 dataset. As detailed in Table 16, we categorize the Top-1G predictions into True Positives and five distinct error types. The comprehensive analysis is as follows:

Significant Gains in True Positives. Our LTKT framework achieves a True Positive (TP) rate of 53.8%, surpassing the baseline (46.4%) by an absolute margin of +7.4%. This indicates that our method fundamentally enhances the model’s ability to retrieve high-quality, continuous action proposals that strictly align with the ground truth.

Mitigation of Background Error. The most notorious challenge in WS-TAL is the false activation on static or visually correlated background frames. Table 16 shows that LTKT drastically reduces the Background Error from 17.4% to 11.8%. This significant drop provides strong quantitative evidence for our Temporal Transition Gating (TTG) mechanism. By strictly enforcing the forward “Preparation → Execution → Completion” causal logic, TTG successfully acts as a physical filter, blocking isolated background noise from generating high-confidence proposals.

Reduction of Localization Error. Localization Error occurs when an action is correctly classified, but its predicted boundaries loosely cover or only partially overlap with the ground truth. LTKT explicitly decreases this error from 31.7% to 28.9%. This improvement stems directly from our Temporal Stage Grounding (TSG) and Linguistic Knowledge Transfer (LKT) modules. Unlike the baseline, which suffers from boundary over-suppression and focuses only on the action climax, our explicit modeling of multi-stage temporal priors ensures that the network accurately captures both the subtle starting motions (preparation) and the ending transitions (completion), yielding much tighter and more precise temporal boundaries.

Table 15: Efficiency and complexity comparison on the THUMOS14 dataset.

Method	GFLOPs	Parameters	Inference Time	AVG mAP
FTCL (Gao et al., 2022)	575	26.5M	36.5ms	43.6
CO2-Net (Hong et al., 2021)	14	34.1M	78.1ms	44.6
UDEL (Chen et al., 2023)	23	52.0M	119.4ms	47.2
MSLU (Shao et al., 2026)	231	28.1M	144.0ms	51.5
Baseline	4.7	6.3M	33.2ms	45.1
LTKT	7.0	10.4M	33.8ms	54.2

Table 16: Proportion of proposal type for Top-1 G prediction results of different models on THUMOS14. G is the number of ground truth instances.

method	Background Err	Localization Err	Double Detection Err	Confusion Err	Wrong Label Err	True Positive
Baseline	17.4%	31.7%	0.3%	3.1%	1.2%	46.4%
LTKT	11.8%	28.9%	0.9%	2.7%	1.9%	53.8%

Analysis of Minor Error Variations. We also observe slight fluctuations in the minor error categories. The Confusion Error sees a marginal decrease (3.1% $\rightarrow$ 2.7%), suggesting that the fine-grained linguistic stage descriptions slightly aid in distinguishing visually similar action categories. Conversely, there is a minor increase in Double Detection Error (0.3% $\rightarrow$ 0.9%) and Wrong Label Error (1.2% $\rightarrow$ 1.9%). This is a mathematically expected phenomenon and a well-known trade-off in detection tasks: because LTKT successfully escapes the severe over-suppression trap of the baseline, it predicts a significantly larger number of valid, continuous proposals (as evidenced by the huge TP gain). With more active foreground predictions, the absolute occurrences of slightly overlapping proposals (Double Detection) naturally experience a slight uptick. Nonetheless, compared to the overwhelming +7.4% gain in True Positives, these minor trade-offs are practically negligible and do not detract from the overall robustness of our framework.

A.6 Details of the Temporal Stage Grounding Prompt

Figure 5 illustrates the exact structured prompt provided to the Large Language Model (LLM) within our Temporal Stage Grounding (TSG) module. The design of this prompt adheres to strict methodological principles to ensure semantic and structural consistency for the generated prior knowledge. First, we define an expert *Role* in video understanding to establish the appropriate semantic domain. Second, we mandate a rigid *Three-Stage Decomposition* (Preparation, Execution, and Completion).

This constraint is crucial, as it explicitly encodes the irreversible chronological and causal flow of human actions into the text embeddings, directly breaking the holistic event fallacy of prior VLM-TAL methods. Furthermore, we provide a concrete *LongJump* example to guide the LLM towards professional descriptions focused on physical body mechanics rather than high-level abstractions, ensuring that the text provides actionable guidance for fine-grained temporal localization.

A.7 Visualization of Feature Distribution

To intuitively understand how our proposed method resolves temporal ambiguity at the representation level, we utilize t-SNE to visualize the snippet-level feature embeddings of two typical action categories (*LongJump* and *CliffDiving*) alongside their corresponding background snippets, as shown in Figure 6.

In the original feature space generated by the pre-trained video backbone, the foreground action features (colored stars) are severely entangled with the background features (grey dots). Conversely, the feature space optimized by our LTKT framework exhibits a remarkable structural transformation. The foreground action features are pulled together into tight, compact clusters, establishing a distinct margin separating them from the background noise. Furthermore, the background features themselves are reorganized into more structured sub-clusters rather than chaotic distributions. This discriminative feature space directly validates the efficacy of our LTKT framework.

Role: You are an expert in Computer Vision and Human Activity Recognition, specializing in temporal action segmentation and video understanding.

Task: I will provide you with a Python list named `CLASS\_LIST` containing action category names. Your goal is to transform this into a new dictionary called `ACTION\_STAGES`.

Requirements:

1. **Three-Stage Decomposition:** For each action, define exactly three continuous and sequential stages:
 - **Stage 1 (Preparation):** The pre-action phase (e.g., wind-up, approach, or initial stance).
 - **Stage 2 (Execution):** The core-action phase (the main movement or the peak of the activity).
 - **Stage 3 (Completion):** The post-action phase (e.g., follow-through, recovery, or returning to a neutral stance).
2. **Output Format:** The output must be a valid Python dictionary.
3. **Language:** Descriptions should be concise, professional, and focus on physical body mechanics.

Example Format: `ACTION_STAGES = {`

```
    'Long Jump': [  
        "The jumper sprints down the runway and takes off from the board.", # Pre-action  
        "The jumper extends the legs forward while soaring through the air.", # Core-action  
        "The jumper lands in the sand pit and moves forward out of the pit." # Post-action    ],  
    }
```

Input: `CLASS_LIST`

Output: Please provide the `ACTION\_STAGES` dictionary now.

Figure 5: The structured prompt provided to the Large Language Model (LLM) for generating multi-stage temporal descriptions. By defining an expert role, enforcing a rigid three-stage physical mechanics constraint, and providing a concrete illustrative example, this prompt ensures consistent, high-quality, and structurally aligned linguistic priors across diverse action categories.

B LLM Usage

During the preparation of this manuscript, AI assistants were utilized exclusively for language polishing and to refine the English writing. The authors confirm that all methodologies, conceptualizations, and scientific findings presented in this paper are entirely original and developed solely by the human authors. AI was not used to generate any core research content, data, or technical concepts. All authors have reviewed the final text and assume full responsibility for the integrity and content of this work.

C Descriptive Statistics

In this paper, we report the model’s performance based on a single run with a fixed random seed. For the inference time in the efficiency experiments, we report the average over five runs.

D Ethics Statement

We utilized the pre-trained I3D (Carreira and Zisserman, 2017) model for feature extraction. The model and its source code are released under the Apache License 2.0, and our use complies with the terms of this license.

We employed the VideoCLIP-XL (Wang et al., 2024) model for cross-modal understanding. This

model is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). Our use of the model in this work is strictly for non-commercial academic purposes, in full compliance with the license terms.

E Parameters for Packages

Our deep learning models were implemented using PyTorch (v1.12) and the transformers library (v4.37) for pre-trained model architectures. Data processing and numerical computations were handled by NumPy (v1.26). For visual tasks, we employed opencv-python (v4.10) for image preprocessing and normalization, while FFmpeg (v4.3) was used for video decoding and frame extraction. Unless otherwise specified, we used the default parameter settings provided by these libraries.

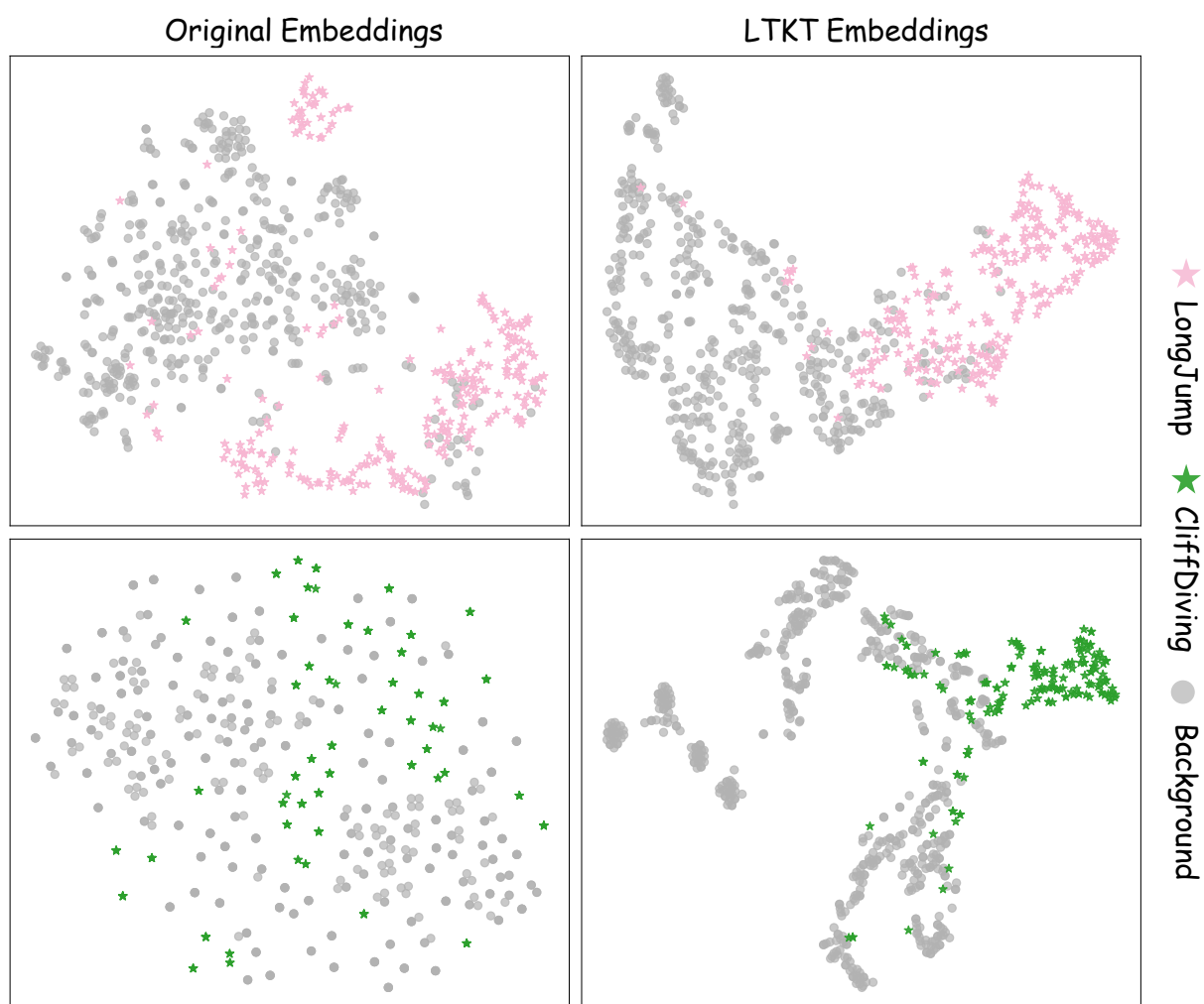


Figure 6: t-SNE visualization of snippet-level feature embeddings for *LongJump* (top) and *CliffDiving* (bottom). **Left:** The original embeddings from the video backbone exhibit severe entanglement between foreground action instances (colored stars) and background snippets (grey dots). **Right:** Embeddings refined by our LTKT framework demonstrate intra-class compactness and inter-class separability, pulling action features away from the ambiguous background.