Summary of Primary Paper: A Survey on Video Moment Localization

**https://arxiv.org/pdf/2306.07515**

# Key Concepts and Taxonomy

The research paper “A Survey on Video Moment Localization” by Meng Liu et al. presents a comprehensive review of the field focused on finding a specific time segment in a video based on a natural language query. This task is also known as video moment retrieval. The paper differentiates this from temporal action localization, which uses predefined action categories, and action recognition, which classifies entire trimmed video clips.

The survey introduces a detailed taxonomy that classifies existing research into three main paradigms: supervised, weakly supervised, and unsupervised learning.

## Supervised Methods

These methods are trained on data with exact start and end time annotations for each video moment corresponding to a query. They are further categorized into:

* + - **Two-Stage Methods**: These first generate potential moment candidates and then match them against the query. The candidate generation can be based on hand-crafted heuristics, multi- scale sliding windows, or specific moment generation networks.
    - **One-Stage Methods**: In these models, the moment generation and localization are integrated into a single, unified framework to improve efficiency. These are classified as anchor-based, sampler-based, or proposal-free, with proposal-free methods directly predicting the start and end times.
    - **Reinforcement Learning (RL) Methods**: These models treat the task as a sequential decision- making problem, where an agent learns a policy to progressively adjust the temporal bound- aries of the target moment.

## Weakly Supervised Methods

To reduce the intensive labor of annotation, these methods are trained using only video-sentence pairs, without the exact temporal boundaries.

## Unsupervised Methods

These methods do not require any textual annotations for the videos during training. They often rely on other resources like visual concept detectors or pre-trained image-sentence embedding spaces.

# Main Contributions and Findings

The paper’s primary contribution is its holistic and fine-grained review of the field, covering work published before December 2021. It provides a more detailed taxonomy than previous surveys and includes a comprehensive comparison of model performance on standard bench- marks like DiDeMo, Charades-STA, TACoS, and ActivityNet-Captions. The survey notes a trend moving from less efficient two-stage methods towards one-stage and RL-based models that offer better speed and an end-to-end approach.

# Limitations and Future Directions

The survey concludes by identifying key challenges and suggesting future research directions.

## Dataset Limitations

Current datasets often have biases, such as moments appearing frequently at the start or end of a video, and queries that are simple and short. The authors call for the creation of large-scale datasets with longer videos and more semantically complex queries.

## Interpretability

A major drawback of current models is their “black box” nature, which makes it hard to un- derstand their reasoning process. The paper suggests that building interpretable models is a promising research direction.

## External Knowledge

The authors propose that augmenting models with external domain knowledge could help them achieve human-level performance, making it an interesting and promising area for future work.

# Key Referenced Papers

## Localizing Moments in Video with Natural Language (Hendricks et al., 2017)

*Key Concepts and Taxonomy*: This paper introduced a two-stage method where videos are first segmented into candidate moments using hand-crafted heuristics. It then learns a shared embedding space for these moments and language queries, using an inter-intra video rank- ing loss to retrieve the best match. Its core model is the Moment Context Network (MCN). *Main Contributions and Findings*: Its main contributions were pioneering the video moment localization task as a retrieval problem and creating the DiDeMo (Distinct Describable Mo- ments) dataset, one of the first and most widely used benchmarks. The model demonstrated the feasibility of localizing moments described by natural language in unedited videos.

*Limitations and Future Directions*: A limitation noted by subsequent research is that MCN considers the entire video as visual context, which can introduce irrelevant information. The survey highlights that later models like MLLC improved upon this by attending to different video contexts conditioned on the query.

*Relevance to Primary Paper*: The survey identifies this as a foundational, “pioneer” paper for two-stage methods and the creator of a key dataset used for comparing many of the models discussed.

## TALL: Temporal Activity Localization via Language Query (Gao et al., 2017)

*Key Concepts and Taxonomy*: This work introduced the Cross-modal Temporal Regression Localizer (CTRL), a model that uses multi-scale temporal sliding windows to generate mo- ment candidates. It processes these candidates and queries through a cross-modal module and uses a multi-task loss for both visual-semantic alignment and temporal boundary re- gression.

*Main Contributions and Findings*: The paper’s key contributions were pioneering the multi- scale sliding window approach for candidate generation and releasing the Charades-STA dataset. The CTRL model became a strong baseline for two-stage methods.

*Limitations and Future Directions*: The survey points out that CTRL’s method of combining video and query information was simple and lacked in-depth analysis of their interaction. Later works like ACRN and ROLE improved upon CTRL by adding attention mechanisms and more sophisticated query modeling.

*Relevance to Primary Paper*: This paper is cited as establishing a key branch of two-stage methods (“pioneer in this branch”) and for providing a second major dataset that is used extensively for evaluation throughout the survey.

## Grounding Action Descriptions in Videos (Regneri et al., 2013)

*Key Concepts and Taxonomy*: This early work focused on grounding action descriptions in procedural videos. It collected data by having crowd-sourced annotators describe the content of video clips with sentences and aligning them to temporal segments.

*Main Contributions and Findings*: The primary contribution of this work was the creation of the TACoS dataset. This dataset features videos of cooking tasks with fine-grained temporal annotations for activities, paired with natural language descriptions.

*Limitations and Future Directions*: As an early dataset, it is limited to a specific domain (cooking). This contrasts with later, open-domain datasets like ActivityNet-Captions.

*Relevance to Primary Paper*: This paper is cited as the source of the TACoS dataset, a challenging and frequently used benchmark in the field due to its long videos and detailed descriptions, which test a model’s ability to handle fine-grained actions.

## Dense-Captioning Events in Videos (Krishna et al., 2017)

*Key Concepts and Taxonomy*: This paper introduced the task of dense video captioning, which involves both localizing and describing multiple events in a single video. To support this, it created a new large-scale dataset.

*Main Contributions and Findings*: Its main contribution was the ActivityNet-Captions dataset. Built on the ActivityNet benchmark, it contains about 20,000 videos and over 70,000 queries, with each video having multiple ground truth segments paired with a caption.

*Limitations and Future Directions*: While a large and valuable resource, the annotations were created for captioning, which can sometimes lead to ambiguity when used for local- ization, a challenge noted in the broader discussion of dataset issues.

*Relevance to Primary Paper*: It is cited as the source of one of the largest and most chal- lenging open-domain datasets for the field. It has become a standard for evaluating the scalability and generalization of video moment localization models.

## Temporally Grounding Natural Sentence in Video (Chen et al., 2018)

*Key Concepts and Taxonomy*: This paper proposed the Temporal GroundNet (TGN), a one- stage, anchor-based model. It sets each time step as an anchor and generates multiple mo- ment candidates ending at that step, allowing for an end-to-end trainable model that is more efficient than two-stage methods.

*Main Contributions and Findings*: Its primary contribution was introducing the first dy- namic, single-stream, end-to-end model for the task. This marked the beginning of the “one- stage” paradigm, which integrates and jointly optimizes candidate generation and matching. *Limitations and Future Directions*: The survey notes that later one-stage methods, such as FIAN, significantly surpassed TGN by incorporating more sophisticated inter-modal inter- action modeling, demonstrating the importance of moving beyond simple feature fusion.

*Relevance to Primary Paper*: This work is highlighted as the “first... end-to-end” model of its kind, marking a major architectural shift in the field focused on improving efficiency and simplifying the training pipeline.

## Span-based Localizing Network for Natural Language Video Localization (Zhang et al., 2020)

*Key Concepts and Taxonomy*: This paper introduced VSLNet, which innovatively framed video moment localization as a span-based question-answering (QA) problem. It is a proposal- free method that uses a context-query attention mechanism to capture cross-modal interac- tions and directly predicts the start and end indices of the “answer” span (the target moment). *Main Contributions and Findings*: Its primary contribution is the novel formulation of the task within a standard QA framework, removing the need for predefined anchors or moment candidates. This approach proved effective, particularly on long videos when extended to its successor, VSLNet-L.

*Limitations and Future Directions*: While effective, the survey notes that proposal-free methods like VSLNet still showed a performance gap compared to top-performing moment generation-based methods at the time. The survey suggests that exploring more effective interaction strategies is a key future direction for these models.

*Relevance to Primary Paper*: The survey highlights VSLNet as a highly influential proposal- free model whose novel QA-based formulation inspired subsequent work and is even used as a component in other models.

## Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language (Zhang et al., 2020)

*Key Concepts and Taxonomy*: This work presented the 2D-TAN model, which represents all possible video moments in a 2D temporal map where axes represent start and end times. By applying a Temporal Adjacent Network (convolutional network) on this 2D map, the model efficiently processes and scores all candidates in a single pass.

*Main Contributions and Findings*: The key contribution is the 2D temporal map represen- tation and the associated network for modeling temporal relations between video moments. This enumeration-based, one-stage method demonstrated strong performance.

*Limitations and Future Directions*: This model was later extended to a multi-scale version (MS-2D-TAN) to better model temporal context at different scales, implying the single-scale version was a limitation.

*Relevance to Primary Paper*: This paper is presented as an influential one-stage method based on enumeration. The survey notes that the 2D map concept is a powerful idea that has been adopted and extended by several subsequent state-of-the-art models.

## Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos (He et al., 2019)

*Key Concepts and Taxonomy*: This paper was the first to formulate video moment local- ization as a sequential decision-making process using an end-to-end reinforcement learning framework, named RWM. An agent observes the video and query, and at each time step,

takes a predefined action (e.g., shift start point, expand right, stop) to progressively adjust a temporal window to match the query.

*Main Contributions and Findings*: Its key contribution was establishing RL as a third major paradigm for supervised localization. This approach avoids exhaustive searches and can lo- calize moments in a few glimpses, making it potentially more efficient.

*Limitations and Future Directions*: The survey notes that while RL methods are efficient, their localization accuracy was inferior to state-of-the-art one-stage methods. The paper suggests this is because they often focus more on policy and reward design while overlook- ing the importance of multimodal fusion and rich context modeling.

*Relevance to Primary Paper*: The survey identifies this as the “first end-to-end reinforce- ment learning based framework”, establishing RL as a viable, intuitive, and efficient ap- proach to solving the video moment localization task.

1. **Weakly Supervised Video Moment Retrieval From Text Queries (Mithun et al., 2019)** *Key Concepts and Taxonomy*: This was the first paper to address video moment localiza- tion under a weakly supervised setting. Its model, TGA (Text-Guided Attention), uses a multiple-instance learning approach where it generates candidate moments and learns a joint embedding, assuming at least one candidate in a positive video corresponds to the query. *Main Contributions and Findings*: The main contribution was demonstrating the feasibility of training a localization model using only video-sentence pairs, significantly reducing the annotation burden. The TGA model served as a de facto baseline for subsequent weakly supervised methods.

*Limitations and Future Directions*: As an early weakly supervised method, later research identified the need for better proposal selection and visual representations. Models like WSLLN and VLANet improved upon TGA by focusing on more robust proposal selection mechanisms to handle spurious candidates.

*Relevance to Primary Paper*: This work is cited as opening up the important and practical research direction of weakly supervised learning for this task (“the first weakly supervised... model”).

1. **Learning Video Moment Retrieval Without a Single Annotated Video (Gao et al., 2021)** *Key Concepts and Taxonomy*: This paper proposed the unpaired video moment retrieval approach (U-VMR), which requires no textual annotations of videos. It leverages exist- ing visual concept detectors and a pre-trained image-sentence embedding space to generate suitable sentence representations for video clips, which are then used for training a moment localizer.

*Main Contributions and Findings*: The key contribution is pioneering the unsupervised paradigm for this task, showing that acceptable localization results can be achieved with- out any paired video-sentence data.

*Limitations and Future Directions*: The performance of unsupervised methods, while im- pressive for the low level of supervision, still lags behind supervised and weakly-supervised approaches. Their performance is also heavily dependent on the quality of the external tools they leverage (e.g., concept detectors, pre-trained embeddings).

*Relevance to Primary Paper*: This paper is cited as the pioneering work in the unsupervised category, pushing the boundary of what is possible with minimal annotation and highlight- ing a key direction for future research in data-scarce domains.

1. **Diagnosing Error in Temporal Action Detectors (Alwassel et al., 2018)**

*Key Concepts and Taxonomy*: This paper introduces the **first diagnostic tool** specifically tailored to analyze error patterns in **temporal action localization**. The authors categorize false positive (FP) errors into five types—Double Detection, Wrong Label, Localization, Confusion, and Background—and analyze their impact on performance. They also characterize dataset instances by six action-specific features: context size, context distance, agreement (between annotators), coverage, length, and number of instances.

*Main Contributions and Findings:*

Developed a comprehensive diagnostic tool for evaluating action detectors on datasets like ActivityNet v1.3.

Proposed a detailed taxonomy of FP errors and quantified their impact using the normalized mAP (average-mAPN).

Found that localization errors had the highest negative impact on performance.

Introduced six new action instance attributes and showed how these characteristics affect detection accuracy.

Demonstrated that temporal agreement between annotators does not significantly hinder performance, validating earlier hypotheses.

*Limitations and Future Directions:*

The analysis is primarily benchmarked on ActivityNet and may need extension to other datasets or tasks.

Future work includes developing evaluation metrics that account for annotation ambiguity and leveraging multiple annotations per instance.

Current insights are empirical; integrating diagnostic feedback into automated modelimprovement pipelines is a potential research direction.

*Relevance to primary paper:*This work is cited as foundational for error analysis in temporal localization. It provides a diagnostic lens that complements performance metrics and has inspired methods targeting localization robustness and contextual reasoning in temporal action models.

1. **Continuum Robots for Medical Applications: A Survey (Burgner-Kahrs et al., 2015)**  
   *Key Concepts and Taxonomy:* This paper presents a comprehensive survey of continuum robots, which are characterized by their continuously bending structures rather than rigid links. These robots are particularly suited for constrained and delicate environments like the human body. The paper categorizes designs based on actuation mechanisms (e.g., cable-driven, concentric tubes, fluidic), kinematics, and control strategies, providing a structured taxonomy of their use in medical procedures.

*Main Contributions and Findings:* The survey establishes continuum robots as a promising technology for minimally invasive medical interventions, offering enhanced dexterity and access in complex anatomical regions. It highlights several clinical prototypes and early deployments, while identifying the versatility of designs for different surgical applications.

*Limitations and Future Directions:* Despite their potential, many continuum robots are still at the research or pre-clinical stage. Challenges include precise real-time shape sensing, dynamic control in soft-tissue environments, and integration with imaging systems. The paper emphasizes the need for miniaturization, robustness, and regulatory approval for clinical adoption.

*Relevance to Primary Paper:* This paper serves as a foundational reference for researchers in medical robotics, especially those focusing on soft and continuum robot design. It identifies the major technical and clinical challenges and lays the groundwork for future innovation in surgical robotics

1. **STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization (Cao et al., 2020)**  
   *Key Concepts and Taxonomy:* This paper formulates video moment localization as a **sequential decision-making** process using reinforcement learning. The proposed STRONG model uses an actor-critic structure to iteratively refine temporal boundaries by incorporating cross-modal representations of video and language.

*Main Contributions and Findings:* STRONG pioneers the use of **reinforcement learning** in this task, enabling better alignment between modalities and fine-tuning of start-end predictions. It significantly outperforms regression-based baselines on standard benchmarks.

*Limitations and Future Directions:* The method still relies on supervised training and could be improved by incorporating weak or unsupervised techniques. Sample efficiency and generalization to longer queries remain open challenges.

*Relevance to Primary Paper:* This paper explores an **alternative formulation** of moment retrieval, offering a distinct path from traditional regression/localization methods and addressing some error modes identified in the diagnostic analysis.

1. **Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization (Cao et al., 2020)**  
   *Key Concepts and Taxonomy:* This paper presents an **adversarial learning framework** that unifies the modeling of ranking relevance and temporal localization. A generator proposes moments, while a discriminator ranks them by semantic alignment with the query.

*Main Contributions and Findings:* By jointly optimizing ranking and localization, the model enhances boundary accuracy and discrimination ability, outperforming existing baselines.

*Limitations and Future Directions:* The approach is still **supervised** and sensitive to annotation quality. Future directions include integrating more robust semantic reasoning or domain adaptation for cross-dataset generalization.

*Relevance to Primary Paper:* This work addresses **localization precision and false positives**, both emphasized in the diagnostic tool from the primary paper. It also strengthens the case for joint modeling strategies.

1. **Temporally Grounding Natural Sentence in Video (Chen et al., 2018)**  
   *Key Concepts and Taxonomy:* One of the earliest works in **temporal sentence grounding**, this paper proposes a regression model that directly predicts the start and end times of moments relevant to a given natural language query.

*Main Contributions and Findings:* Demonstrates that end-to-end regression over visual-linguistic features is feasible and effective. Serves as a baseline for later methods in moment localization.

*Limitations and Future Directions:* The model lacks **explicit context modeling** and struggles with ambiguous queries. Future research could integrate semantic structures or use multi-stage reasoning.

*Relevance to Primary Paper:* As a foundational supervised method, it provides a benchmark for evaluating **temporal localization performance and error types** diagnosed in the primary study.

1. **Localizing Natural Language in Videos (Chen et al., 2019)**  
   *Key Concepts and Taxonomy:* This work introduces a **proposal-based framework** for sentence grounding in videos, where candidate segments are generated and scored using attention-based cross-modal embeddings.

*Main Contributions and Findings:* Achieves better boundary precision and robustness than direct regression models. The architecture is modular and easier to analyze and optimize.

*Limitations and Future Directions:* Performance is limited by proposal quality and alignment noise. Future improvements may include **dynamic proposal adjustment** or end-to-end differentiable alternatives.

*Relevance to Primary Paper:* Addresses **localization and background errors**, two core areas identified by the diagnostic tool in the primary paper, and marks a shift toward interpretable, modular grounding.

1. **Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos (Shaoxiang Chen et al., 2020)**  
   *Key Concepts and Taxonomy:* A **multi-task learning framework** that jointly models sentence localization and video captioning by leveraging cross-modal interactions.

*Main Contributions and Findings:* Shows that shared representation learning across modalities improves both grounding and captioning. Incorporates attention mechanisms to align video-sentence features.

*Limitations and Future Directions:* Requires careful balancing between tasks and is sensitive to noisy cross-modal representations. Future work could explore adaptive multi-task weighting or modality disentanglement.

*Relevance to Primary Paper:* This model supports the **integration of multiple tasks** to enhance localization robustness, complementing the diagnostic paper’s emphasis on action characteristics and representation learning.

1. **Semantic Proposal for Activity Localization in Videos via Sentence Query (Shaoxiang Chen & Yugang Jiang, 2019)**  
   *Key Concepts and Taxonomy:* Proposes a **semantic proposal generation** mechanism that filters video segments based on their alignment with sentence semantics before final localization.

*Main Contributions and Findings:* Improves precision by eliminating irrelevant proposals early. The method reduces false positives and enhances localization accuracy, especially for complex or long videos.

*Limitations and Future Directions:* Performance may drop for unseen sentence semantics or loosely aligned queries. Future research could enhance generalization using **semantic expansion or retrieval-augmented models**.

*Relevance to Primary Paper:* Tackles two central issues from the primary paper: **semantic confusion** and **false positives**, by bringing linguistic understanding into the proposal stage.

1. **End-to-End Multi-Modal Video Temporal Grounding (Yi-Wen Chen et al., 2021)**  
   *Key Concepts and Taxonomy:* This paper proposes an **end-to-end learning framework** for temporal grounding of natural language in videos, utilizing a unified architecture that jointly processes multi-modal inputs without relying on pre-generated proposals or post-processing.

*Main Contributions and Findings:* The model directly predicts start and end times of relevant video segments from sentence queries by fusing visual and textual features through multi-modal attention mechanisms. It achieves **state-of-the-art performance** on several benchmarks including Charades-STA and TACoS.

*Limitations and Future Directions:* Although efficient and accurate, the model may be limited by **dataset-specific tuning** and could benefit from better generalization across video domains and query styles. Future work could explore adaptation to zero-shot settings or longer video sequences.

*Relevance to Primary Paper:* This work exemplifies the **proposal-free supervised** category of moment localization approaches and directly addresses the boundary prediction accuracy issue highlighted by the diagnostic tool in the primary paper.

1. **Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video (Zhenfang Chen et al., 2019)**  
   *Key Concepts and Taxonomy:* This paper introduces a **weakly-supervised** framework to temporally and spatially ground natural language sentences in video without detailed moment-level supervision. It combines visual and language embeddings in a joint space and applies attention mechanisms to highlight relevant regions over time.

*Main Contributions and Findings:* The approach proves that spatio-temporal grounding is possible with only coarse video-level annotations. It introduces techniques to refine proposals through contrastive learning and localization-aware alignment.

*Limitations and Future Directions:* Localization accuracy remains lower than fully supervised methods. Future work could address proposal quality and explore self-supervised pretraining to boost visual grounding performance.

*Relevance to Primary Paper:* This method expands weakly-supervised video moment retrieval to include **spatial grounding**, contributing to the understanding of localization errors across dimensions, which the primary paper categorizes and diagnoses.

1. **A Review of Motion Planning for Highway Autonomous Driving (Claussmann et al., 2020)**  
   *Key Concepts and Taxonomy:* This review surveys **motion planning strategies** for autonomous driving on highways, organizing them into rule-based, optimization-based, and learning-based methods, and analyzing their performance under different constraints like safety, comfort, and computational cost.

*Main Contributions and Findings:* It presents a **comparative study** of real-time feasibility, environment interaction, and scalability for each method, identifying current gaps in robustness and adaptability under uncertainty.

*Limitations and Future Directions:* The field lacks standardized evaluation benchmarks, and learning-based planners require further advances in **safe generalization** and explainability.

*Relevance to Primary Paper:* While not directly related to video moment localization, this survey highlights **trajectory-level decision problems**, relevant to the design of agents for navigation tasks in video understanding and simulation contexts.

1. **Temporal Localization of Moments in Video Collections with Natural Language (Escorcia et al., 2019)**  
   *Key Concepts and Taxonomy:* This paper explores **video-level retrieval** across large-scale collections, introducing a two-step pipeline: first retrieving candidate videos, then localizing moments within them using natural language queries.

*Main Contributions and Findings:* Proposes the concept of **coarse-to-fine moment retrieval**, demonstrating strong performance on datasets like DiDeMo. It uses visual-language alignment scores at both levels to prune search space and refine localization.

*Limitations and Future Directions:* Performance is dependent on candidate retrieval quality. Future directions could integrate **dense video indexing** and multilingual query handling.

*Relevance to Primary Paper:* This work contributes to identifying **retrieval and localization failures**, especially false negatives caused by retrieval bias—a key category discussed in the primary paper.

1. **TALL: Temporal Activity Localization via Language Query (Jiyang Gao et al., 2017)**  
   *Key Concepts and Taxonomy:* TALL is among the **first end-to-end supervised** models for grounding language in untrimmed video streams. It uses a Cross-modal Temporal Regression Localizer (CTRL) to map sentence and video pairs to temporal coordinates.

*Main Contributions and Findings:* Introduces the idea of **joint feature embedding and regression** for moment boundaries. It provides baseline metrics for popular datasets like Charades-STA and TACoS.

*Limitations and Future Directions:* Limited scalability and precision due to reliance on fixed-length features and handcrafted pooling. Future work could adopt **adaptive attention** and multi-resolution analysis.

*Relevance to Primary Paper:* TALL is a foundational benchmark for **supervised models**, referenced in the primary paper's analysis of regression-based localization errors.

1. **Relation-aware Video Reading Comprehension for Temporal Language Grounding (Jialin Gao et al., 2021)**  
   *Key Concepts and Taxonomy:* This paper treats moment localization as a **video question-answering task**, introducing a relation-aware model that captures interactions between objects, actions, and scenes using graph networks.

*Main Contributions and Findings:* Proposes a **video reading comprehension paradigm** that enhances temporal grounding by leveraging spatio-temporal and semantic relations. Achieves competitive results on multiple datasets.

*Limitations and Future Directions:* Model complexity increases training cost. Performance could degrade under **weak supervision** or noisy input. Future work may reduce dependency on external parsers or expand to multi-modal QA.

*Relevance to Primary Paper:* This method exemplifies the importance of **semantic reasoning and relational understanding**, both key areas suggested for error mitigation in the diagnostic framework.

1. **Fast Video Moment Retrieval (Junyu Gao and Changsheng Xu, 2021)**  
   *Key Concepts and Taxonomy:* This work proposes a **lightweight and efficient framework** for video moment retrieval, focusing on speeding up inference and reducing computational cost without significantly sacrificing accuracy.

*Main Contributions and Findings:* The method introduces fast segment matching using **dual attention mechanisms** and optimized segment sampling. It achieves competitive accuracy with **much lower latency** than traditional models.

*Limitations and Future Directions:* While efficient, the model may lack semantic depth for complex queries. Future work could enhance **semantic reasoning** or fuse external knowledge for harder localization scenarios.

*Relevance to Primary Paper:* This method addresses the **computational efficiency and scalability challenge**, a practical aspect relevant to deploying systems beyond lab settings—a limitation acknowledged in the primary paper.

1. **Learning Video Moment Retrieval Without a Single Annotated Video (Junyu Gao and Changsheng Xu, 2021)**  
   *Key Concepts and Taxonomy:* This paper proposes an **unsupervised training strategy** for moment retrieval using pseudo-labels generated via video-text alignment without any ground-truth moment annotations.

*Main Contributions and Findings:* It introduces a **cycle-consistency mechanism** to train the retrieval model from raw video-text pairs. Despite the lack of annotations, it performs competitively against weakly supervised baselines.

*Limitations and Future Directions:* Performance remains behind supervised methods, especially in complex language scenarios. Future work could improve **pseudo-label refinement** or incorporate minimal supervision.

*Relevance to Primary Paper:* This work aligns with the push toward **reducing annotation cost**, directly responding to the scalability challenges and supervision limitations highlighted in the primary diagnostic framework.

1. **WSLLN: Weakly Supervised Natural Language Localization Networks (Mingfei Gao et al., 2019)**  
   *Key Concepts and Taxonomy:* WSLLN introduces a **weakly supervised** model that aligns sentences and video segments without temporal labels. It uses **multiple instance learning** and frame-level relevance prediction to localize moments.

*Main Contributions and Findings:* It learns a joint embedding space and applies max-margin learning to differentiate relevant and irrelevant video parts, establishing a strong baseline for weak supervision.

*Limitations and Future Directions:* Susceptible to spurious alignment during training. Future models could benefit from **semantic priors or attention regularization**.

*Relevance to Primary Paper:* WSLLN is among the first to improve upon early weakly supervised models like TGA, addressing the **proposal quality and alignment issues** raised in the diagnostic analysis.

1. **MAC: Mining Activity Concepts for Language-Based Temporal Localization (Runzhou Ge et al., 2019)**  
   *Key Concepts and Taxonomy:* MAC proposes to improve localization by **explicitly mining activity concepts** (verbs and objects) from language queries and aligning them with visual features to improve grounding precision.

*Main Contributions and Findings:* The model uses a dual attention mechanism over both concept and sentence-level embeddings, demonstrating improved accuracy in understanding fine-grained actions.

*Limitations and Future Directions:* Relies on pre-defined concept extraction tools, which may limit generalization. Future directions could include **end-to-end concept discovery** or multimodal pretraining.

*Relevance to Primary Paper:* This method tackles **semantic confusion errors**, offering an effective solution pathway that complements the findings in the error categorization framework.

1. **ExCL: Extractive Clip Localization Using Natural Language Descriptions (Soham Ghosh et al., 2019)**  
   *Key Concepts and Taxonomy:* ExCL frames moment retrieval as a **sequence tagging problem**, where each video frame is labeled to determine inclusion in the relevant clip. It applies a transformer-based encoder to learn global context.

*Main Contributions and Findings:* Demonstrates that **fine-grained frame-level alignment** improves performance, especially on long and untrimmed videos. Achieves state-of-the-art results on ActivityNet Captions.

*Limitations and Future Directions:* High memory cost due to frame-wise tagging. Could benefit from hierarchical modeling or sparsity constraints to scale better.

*Relevance to Primary Paper:* This method directly addresses **boundary and background errors**, two primary failure types discussed in the diagnostic paper.

1. **Tripping Through Time: Efficient Localization of Activities in Videos (Meera Hahn et al., 2020)**  
   *Key Concepts and Taxonomy:* This work introduces an efficient model for **localizing action moments** in video streams using temporal attention and clip-level embeddings. It targets fast inference and high accuracy by segmenting videos into candidate clips and filtering them using relevance scores.

*Main Contributions and Findings:* Demonstrates a lightweight but effective mechanism to **scale localization** in longer videos. The approach outperforms several prior works on ActivityNet and THUMOS datasets in both speed and accuracy.

*Limitations and Future Directions:* Focuses primarily on speed; lacks deep semantic reasoning. Future extensions could integrate **contextual understanding or transformer-based modeling**.

*Relevance to Primary Paper:* Directly addresses **boundary and background errors**, as highlighted in the diagnostic study, and supports the trend of balancing performance and scalability.

1. **Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos (Dongliang He et al., 2019)**  
   *Key Concepts and Taxonomy:* This paper models moment localization as a **sequential decision problem** using reinforcement learning. The agent reads a sentence, watches a video, and iteratively adjusts predicted segment boundaries.

*Main Contributions and Findings:* Outperforms traditional regression methods by learning better boundary refinement policies. Introduces a reward signal based on IoU improvements and temporal continuity.

*Limitations and Future Directions:* Training can be unstable due to **sparse rewards**. Future work may explore hierarchical or curriculum reinforcement learning to stabilize convergence.

*Relevance to Primary Paper:* Aligns with the diagnostic paper’s call for **iterative and context-aware models** to mitigate localization drift and improve precision.

1. **Localizing Moments in Video with Temporal Language (Lisa Anne Hendricks et al., 2018)**  
   *Key Concepts and Taxonomy:* This work addresses temporal grounding using **language-conditioned classifiers** trained to score temporal intervals based on alignment with sentence queries. Proposes a dataset and evaluation protocol.

*Main Contributions and Findings:* Demonstrates the feasibility of learning direct mappings between language and time intervals. Introduced a key benchmark dataset (DiDeMo) for moment localization research.

*Limitations and Future Directions:* Fixed sampling rate and uniform proposals limit fine-grained precision. Future directions include adaptive sampling or **multimodal fusion with richer context**.

*Relevance to Primary Paper:* Provides baseline data and error types (e.g., false positives, ambiguous intervals) that informed the categorization strategy used in the primary diagnostic tool.

1. **Natural Language Object Retrieval (Ronghang Hu et al., 2016)**  
   *Key Concepts and Taxonomy:* Focuses on grounding natural language in static images (rather than video) by retrieving object regions described by phrases using spatial attention and multimodal fusion.

*Main Contributions and Findings:* Introduces a joint vision-language embedding with attention for spatial localization. Forms the basis for follow-up work in temporal and spatio-temporal localization.

*Limitations and Future Directions:* Limited to static visual grounding. Future work (which indeed followed) extended the method to video and temporal domains.

*Relevance to Primary Paper:* Influences **spatio-temporal alignment models** and proposal filtering strategies, relevant in the broader context of reducing visual-semantic mismatches highlighted in the primary work.

1. **Video Moment Localization via Deep Cross-Modal Hashing (Yupeng Hu et al., 2021)**  
   *Key Concepts and Taxonomy:* Introduces a **cross-modal hashing approach** for moment retrieval, enabling efficient retrieval through compact binary codes learned via deep supervision.

*Main Contributions and Findings:* Proposes a contrastive loss to align sentence-video segment pairs in the Hamming space. Reduces retrieval time significantly while maintaining competitive accuracy.

*Limitations and Future Directions:* Binary codes may lose fine-grained temporal precision. Future work could explore **hybrid hashing-embedding techniques**.

*Relevance to Primary Paper:* Contributes to **retrieval-stage error analysis**, showing how fast retrieval methods could affect false negatives or coarse localization, both of which are discussed in the diagnostic framework.

1. **Coarse-to-Fine Semantic Alignment for Cross-Modal Moment Localization (Yupeng Hu et al., 2021)**  
   *Key Concepts and Taxonomy:* This paper introduces a **coarse-to-fine framework** for aligning natural language descriptions with video moments. It uses a two-stage architecture: a coarse alignment module for rough localization and a fine module for refining temporal boundaries.

*Main Contributions and Findings:* Achieves strong performance by integrating **hierarchical attention mechanisms** and dual alignment strategies. Demonstrates state-of-the-art results on TACoS and Charades-STA datasets.

*Limitations and Future Directions:* The two-stage nature increases computation. Future work might explore **end-to-end fusion** or improve the robustness of the coarse proposal stage.

*Relevance to Primary Paper:* Directly contributes to solutions addressing **boundary prediction and localization drift**, both of which are major diagnostic categories in the primary paper.

1. **Cross-Modal Video Moment Retrieval with Spatial and Language-Temporal Attention (Bin Jiang et al., 2019)**  
   *Key Concepts and Taxonomy:* Proposes an attention-based model that captures both **spatial cues** in video and **temporal alignment** with language for precise moment retrieval. The model fuses spatial features with language embeddings via dual-attention pathways.

*Main Contributions and Findings:* Shows the importance of joint spatial-temporal attention and achieves notable improvements in retrieval accuracy. Supports better context modeling between actions and objects.

*Limitations and Future Directions:* Heavier reliance on spatial feature extraction may hinder performance on low-quality or occluded videos. Future work could address robustness and **multi-scale temporal fusion**.

*Relevance to Primary Paper:* Helps mitigate **visual-semantic alignment errors** and localization imprecision, core challenges diagnosed in the error taxonomy.

1. **Skip-Thought Vectors (Ryan Kiros et al., 2015)**  
   *Key Concepts and Taxonomy:* This foundational NLP work introduces an **unsupervised sentence encoder**, Skip-Thought, which predicts surrounding sentences from a given one using an RNN-based sequence-to-sequence model.

*Main Contributions and Findings:* Shows that sentence-level representations learned in an unsupervised way generalize well across multiple tasks. The model set the stage for subsequent contextualized language models.

*Limitations and Future Directions:* Skip-Thought lacks fine-tuning capability and is outperformed by modern transformers. Future directions led to **BERT-style models** with superior contextual understanding.

*Relevance to Primary Paper:* The model is frequently used in **video-language retrieval baselines**, serving as the text encoder in earlier video moment localization pipelines assessed in the diagnostic study.

1. **Dense-Captioning Events in Videos (Ranjay Krishna et al., 2017)**  
   *Key Concepts and Taxonomy:* This paper addresses the joint task of **temporal event detection and language captioning** in videos. It uses proposal generation followed by caption generation conditioned on video clips.

*Main Contributions and Findings:* Introduced a new dataset and benchmark. Proposes a recurrent captioning model for densely predicted event segments. Groundbreaking in linking localization and generation.

*Limitations and Future Directions:* Event boundaries are still noisy; future improvements could involve **better proposal refinement or multimodal training**.

*Relevance to Primary Paper:* Provides groundwork for analyzing **language-conditioned proposals**, crucial for understanding **false positives from mismatched captions** as detailed in the diagnostic tool.

1. **Bidirectional Single-Stream Temporal Sentence Query Localization in Untrimmed Videos (Cheng Li et al., 2019)**  
   *Key Concepts and Taxonomy:* Presents a **single-stream model** that simultaneously learns to regress the start and end points of moments using bidirectional modeling of the video sequence.

*Main Contributions and Findings:* Improves localization accuracy by considering **bidirectional temporal context** rather than independent boundaries. Introduces a contrastive loss for better discriminative learning.

*Limitations and Future Directions:* Struggles with long or repetitive actions. Future extensions could involve **temporal abstraction layers** or memory-augmented reasoning.

*Relevance to Primary Paper:* The bidirectional modeling strategy aligns with the paper’s call for **better boundary regression models**, especially those that incorporate global context for reducing localization errors.

1. **Multi-scale 2D Representation Learning for Weakly-supervised Moment Retrieval (Ding Li et al., 2021)**  
   *Key Concepts and Taxonomy:* This paper presents a **2D convolutional model** that processes moment-sentence pairs as a 2D matrix of cross-modal similarities. It emphasizes learning representations at multiple temporal resolutions for better localization under weak supervision.

*Main Contributions and Findings:* Introduces a **multi-scale feature extractor** that allows the model to capture fine-grained and long-range temporal dependencies. Delivers improved results over existing weakly-supervised baselines on Charades-STA and ActivityNet Captions.

*Limitations and Future Directions:* The method may suffer from imprecision due to fixed window structures. Future work could integrate **dynamic attention** or contrastive temporal learning.

*Relevance to Primary Paper:* Supports the diagnostic framework’s call for **stronger proposal modeling** and better context integration in weakly-supervised settings.

1. **Global-local Temporal Representations for Video Person Re-identification (Jianing Li et al., 2019)**  
   *Key Concepts and Taxonomy:* Though centered on **person re-identification**, this paper introduces temporal representation strategies combining **global and local temporal cues**, which are relevant for sequence-level video understanding.

*Main Contributions and Findings:* Employs a two-branch network to extract **global context** and **segment-level features**, improving accuracy in matching individuals across videos under varying conditions.

*Limitations and Future Directions:* Task-specific design limits direct application to general moment retrieval. Could be adapted for **video grounding tasks** with suitable alignment mechanisms.

*Relevance to Primary Paper:* Offers transferable insights for **temporal feature aggregation**, which is crucial in reducing background and confusion errors in video grounding as analyzed in the diagnostic tool.

1. **Multi-scale Temporal Cues Learning for Video Person Reidentification (Jianing Li et al., 2020)**  
   *Key Concepts and Taxonomy:* This work builds on the previous model with a **multi-scale temporal cues learning** architecture, designed to address scale variance in person re-identification from video streams.

*Main Contributions and Findings:* Integrates multi-scale convolutional and recurrent layers to capture both **short- and long-term dependencies**, achieving high accuracy across several re-ID benchmarks.

*Limitations and Future Directions:* Model complexity can affect inference time. Future exploration could involve **temporal attention compression** or unifying the model for other sequence tasks like moment retrieval.

*Relevance to Primary Paper:* Contributes **generalizable temporal modeling strategies** that address issues such as misaligned or incomplete predictions discussed in the primary paper's error taxonomy.

**43. Proposal-Free Video Grounding with Contextual Pyramid Network (Kun Li et al., 2021)**  
*Key Concepts and Taxonomy:* This paper proposes a **proposal-free** method for temporal sentence grounding. It introduces a Contextual Pyramid Network (CPNet), which eliminates the need for pre-generated moment proposals by learning to predict grounding boundaries directly.

*Main Contributions and Findings:* CPNet builds a multi-scale contextual pyramid to capture hierarchical features across different time scales. This significantly improves grounding accuracy while reducing computation.

*Limitations and Future Directions:* Being proposal-free simplifies training but may lack fine-grained control over temporal boundaries. Future extensions may incorporate **learned confidence-based interval refinement**.

*Relevance to Primary Paper:* Aligns with the primary paper’s goal of **better proposal quality** and contextual reasoning, offering a new angle through end-to-end grounding without explicit proposal generation.

**44. Local-enhanced Interaction for Temporal Moment Localization (Guoqiang Liang et al., 2021)**  
*Key Concepts and Taxonomy:* This work focuses on improving local interactions between query and video frames using a Local-Enhanced Interaction Module (LEIM) for better moment localization.

*Main Contributions and Findings:* The paper enhances fine-grained alignment by enriching local video-sentence correspondence, achieving strong results on standard benchmarks.

*Limitations and Future Directions:* While effective locally, it may miss broader context needed for long-range dependencies. The authors suggest integrating global modeling techniques.

*Relevance to Primary Paper:* Supports the **fine-grained visual-linguistic alignment** needed to reduce confusion and background errors as diagnosed in the primary study.

**45. Continuous Control with Deep Reinforcement Learning (Lillicrap et al., 2016)**  
*Key Concepts and Taxonomy:* A foundational paper introducing **Deep Deterministic Policy Gradient (DDPG)** for continuous control tasks in reinforcement learning.

*Main Contributions and Findings:* The key innovation lies in combining deterministic policies with actor-critic methods using deep networks. Though not directly tied to video grounding, its ideas influence **reinforcement-based grounding approaches**.

*Limitations and Future Directions:* General-purpose RL paper; requires careful tuning and suffers from instability in high-dimensional video tasks.

*Relevance to Primary Paper:* Not directly cited, but foundational for **reinforcement-based grounding models** like “Read, Watch, and Move” that address sequential temporal decisions.

**46. Single Shot Temporal Action Detection (Tianwei Lin et al., 2017)**  
*Key Concepts and Taxonomy:* Introduces SS-TAD, a single-stage model for temporal action detection, inspired by the SSD object detector. It treats videos as sequences of frames and predicts start-end times and categories in one pass.

*Main Contributions and Findings:* Reduces latency by avoiding proposal generation. Useful for real-time scenarios. Strong results on THUMOS14 and ActivityNet.

*Limitations and Future Directions:* Lacks rich temporal reasoning. Future work could explore **context-aware feature augmentation** or **multi-granularity modeling**.

*Relevance to Primary Paper:* SS-TAD's one-shot approach is a baseline for **proposal quality** and temporal precision—key error sources discussed in the paper.

**47. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network (Zhijie Lin et al., 2020)**  
*Key Concepts and Taxonomy:* Proposes a Semantic Completion Network (SCNet) under weak supervision, which fills in semantic gaps in candidate proposals using sentence-level constraints.

*Main Contributions and Findings:* Introduces semantic completion loss to ensure proposal-level semantics align with the sentence, outperforming other weakly supervised baselines.

*Limitations and Future Directions:* Quality depends on how well semantic cues generalize across domains. Future work may integrate **external knowledge graphs or commonsense priors**.

*Relevance to Primary Paper:* Enhances **proposal-sentence alignment**, directly addressing a diagnostic concern of weakly supervised retrieval raised in the primary paper.

**48. Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction (Z. Lin et al., 2020)**  
*Key Concepts and Taxonomy:* This paper presents a Cross-Modal Interaction Network (CMIN) that aligns video and textual modalities while reconstructing the query to improve grounding precision.

*Main Contributions and Findings:* CMIN improves temporal grounding by enforcing semantic consistency between the localized moment and the original query through reconstruction loss. It uses mutual enhancement between video and language features.

*Limitations and Future Directions:* While effective, the reconstruction approach may struggle with abstract or ambiguous queries. Future research could focus on incorporating external commonsense or discourse-level reasoning.

*Relevance to Primary Paper:* The work supports the diagnostic insights that highlight **semantic misalignment** as a critical error type in moment retrieval, offering a concrete architectural enhancement to mitigate it.

**49. Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos (Bingbin Liu et al., 2018)**  
*Key Concepts and Taxonomy:* Introduces a **modular retrieval network** that decomposes complex sentences and matches them with corresponding sub-activities in the video.

*Main Contributions and Findings:* The model dynamically assembles modular components to match temporal compositions in long videos. It excels at grounding multi-step tasks described in compound queries.

*Limitations and Future Directions:* The approach depends heavily on accurate query decomposition. More robust semantic parsing and modular selection mechanisms could improve performance.

*Relevance to Primary Paper:* This model addresses limitations in **compositional reasoning**, an important direction noted in the primary paper’s failure analysis for handling complex, long-range queries.

**50. Reasoning Step-by-Step: Temporal Sentence Localization in Videos via Deep Rectification-Modulation Network (Daizong Liu et al., 2020)**  
*Key Concepts and Taxonomy:* Proposes a Deep Rectification-Modulation Network (RMN) that models **step-by-step temporal reasoning** for localizing sentences in video.

*Main Contributions and Findings:* The rectification module refines proposals by suppressing distractors, while the modulation module adjusts features to better fit the query semantics.

*Limitations and Future Directions:* RMN improves proposal refinement but may still face issues with vague queries or background clutter. It could benefit from joint optimization with global context models.

*Relevance to Primary Paper:* RMN directly addresses **background confusion and weak proposal filtering**, two of the key error types highlighted in the diagnostic evaluation.

**51. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos (Daizong Liu et al., 2021)**  
*Key Concepts and Taxonomy:* Proposes an adaptive proposal generation network that customizes candidate segments according to query semantics using cross-modal interaction.

*Main Contributions and Findings:* It dynamically adapts anchor proposals based on sentence features, improving localization precision. Outperforms static proposal models.

*Limitations and Future Directions:* The model relies on pre-trained embeddings and can be less effective with noisy sentence inputs. Future work could explore **robust joint training** on sentence-video pairs.

*Relevance to Primary Paper:* Contributes toward **proposal quality refinement**, a critical axis in reducing localization errors identified in the primary paper.

**52. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding (Daizong Liu et al., 2021)**  
*Key Concepts and Taxonomy:* Introduces a Biaffine Localizing Network (BLN) that models bidirectional interactions between sentence and video with context-aware scoring for start and end positions.

*Main Contributions and Findings:* Uses biaffine transformation to simultaneously predict start-end boundaries and captures rich contextual dependencies. Achieves state-of-the-art accuracy.

*Limitations and Future Directions:* Performance may degrade on datasets with extreme variability in action lengths. Integrating multi-scale temporal modeling could help generalize better.

*Relevance to Primary Paper:* Tackles **start/end mislocalization** errors by learning fine-grained boundary estimation, echoing the diagnostic emphasis on precision timing.

**53. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization (Daizong Liu et al., 2020)**  
*Key Concepts and Taxonomy:* Introduces a **graph-based model** for temporal grounding, which models both **cross-modal (sentence-video)** and **self-modal (intra-video and intra-sentence)** relationships via Graph Attention Networks (GATs).

*Main Contributions and Findings:* The joint modeling of intra- and inter-modal dependencies leads to better semantic alignment and improved moment localization. Shows gains over prior non-graph models.

*Limitations and Future Directions:* Graph-based models can be **computationally intensive**, especially on longer videos. Future work could focus on **sparse graph strategies or transformer-based alternatives**.

*Relevance to Primary Paper:* Directly addresses **semantic misalignment and proposal quality issues**, aligning well with the error categories discussed in the diagnostic evaluation.

**54. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding (Daizong Liu et al., 2021)**  
*Key Concepts and Taxonomy:* Proposes an **iterative refinement framework** that gradually aligns sentence and video representations for better grounding accuracy.

*Main Contributions and Findings:* It introduces a progressively guided attention mechanism, allowing the model to iteratively correct its localization by focusing on relevant parts of the video.

*Limitations and Future Directions:* May struggle with ambiguous or noisy queries where progressive refinement is not meaningful. Future work could integrate **query disambiguation modules**.

*Relevance to Primary Paper:* Strongly contributes to solving **localization drift and boundary imprecision**, aligning with key diagnostic failure modes.

**55. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning (M. Liu et al., 2019)**  
*Key Concepts and Taxonomy:* Focuses on **categorizing short-form (micro) videos** using structure-guided multimodal features through dictionary learning.

*Main Contributions and Findings:* Offers a compact, interpretable representation for video classification using joint visual and textual patterns. Useful for retrieval and tagging in video platforms.

*Limitations and Future Directions:* Less applicable to long-form or untrimmed video grounding tasks. Could be extended with **temporal modeling or contextual embeddings**.

*Relevance to Primary Paper:* While not directly related to moment localization, it informs **multimodal feature alignment**—a supporting component in temporal grounding pipelines.

**56. Iterative Local-Global Collaboration Learning Towards One-Shot Video Person Re-Identification (M. Liu et al., 2020)**  
*Key Concepts and Taxonomy:* A **one-shot learning approach** to video-based person re-ID, combining local appearance and global motion features through an iterative collaboration mechanism.

*Main Contributions and Findings:* Effectively fuses fine-grained spatial and coarse temporal information, achieving strong results with minimal labeled data.

*Limitations and Future Directions:* Not focused on sentence-based localization, but techniques could be adapted for **temporal feature fusion** in grounding models.

*Relevance to Primary Paper:* Indirectly relevant—contributes ideas to **temporal and multi-scale feature fusion**, important in segment-level video understanding.

**57. Attentive Moment Retrieval in Videos (Meng Liu et al., 2018)**  
*Key Concepts and Taxonomy:* Proposes a model that aligns sentence and video features via **attention-based mechanisms**, performing retrieval over sliding windows in the video.

*Main Contributions and Findings:* Introduces a two-stage process of candidate proposal generation and attention-based matching. Strong early baseline for cross-modal retrieval.

*Limitations and Future Directions:* Uses a fixed proposal strategy, which can limit flexibility. Future work may benefit from **adaptive or end-to-end proposal refinement**.

*Relevance to Primary Paper:* Cited as an early attention-based baseline, contributing foundational insights on **sentence-conditioned temporal scoring**.

**58. Cross-Modal Moment Localization in Videos (Meng Liu et al., 2018)**  
*Key Concepts and Taxonomy:* Introduces a framework for **cross-modal retrieval**, aligning sentence and video embeddings to localize relevant video moments.

*Main Contributions and Findings:* Employs a bi-directional ranking loss for alignment and uses semantic matching across modalities. The model operates in two stages: proposal generation and ranking-based retrieval.

*Limitations and Future Directions:* Struggles with vague queries and fixed proposals. Future work could improve proposal flexibility and incorporate **contextual alignment techniques**.

*Relevance to Primary Paper:* Forms one of the early **ranking-based baselines**, tackling core challenges around semantic matching that are deeply analyzed in the diagnostic evaluation.

**59. SSD: Single Shot MultiBox Detector (Wei Liu et al., 2016)**  
*Key Concepts and Taxonomy:* SSD is a **one-stage object detector** designed for real-time object detection by predicting bounding boxes and class scores in a single pass.

*Main Contributions and Findings:* Introduces default anchor boxes with different aspect ratios and scales, offering high accuracy and speed. Influences temporal analogs like **SS-TAD** in video action detection.

*Limitations and Future Directions:* Initially developed for images; limited temporal reasoning. Later works adapt its ideas to action localization.

*Relevance to Primary Paper:* Inspired models like **SS-TAD** that follow a similar detection strategy in the **temporal domain**, affecting proposal design quality.

**60. A Survey on Natural Language Video Localization (Xinfang Liu et al., 2021)**  
*Key Concepts and Taxonomy:* This is a **survey paper** reviewing developments in natural language-based video moment localization, covering both supervised and weakly-supervised approaches.

*Main Contributions and Findings:* Provides a comprehensive taxonomy, compares benchmark results, and discusses challenges such as ambiguous queries, weak supervision, and proposal design.

*Limitations and Future Directions:* Lacks in-depth discussion on **error analysis**. Future surveys may integrate tools like DETAD for evaluation and interpretability.

*Relevance to Primary Paper:* Complements the diagnostic study by summarizing the **landscape of methods**, many of which are analyzed in the main paper.

**61. Single-shot Semantic Matching Network for Moment Localization in Videos (Xinfang Liu et al., 2021)**  
*Key Concepts and Taxonomy:* Proposes a **single-shot localization model** that uses semantic matching between the query and video frames without explicit proposal generation.

*Main Contributions and Findings:* Achieves fast and accurate retrieval through **end-to-end sentence-conditioned predictions**. It handles both localization and matching simultaneously.

*Limitations and Future Directions:* May lack interpretability and struggle with long-term dependencies. Potential enhancements include hierarchical temporal modeling.

*Relevance to Primary Paper:* Directly contributes to solving the **proposal quality bottleneck** through a proposal-free strategy, addressing one of the core diagnostic axes.

**62. Debug: Dense Bottom-Up Grounding Approach for Natural Language Video Localization (Chujie Lu et al., 2019)**  
*Key Concepts and Taxonomy:* Presents a **dense, bottom-up approach** to localize moments by scoring every frame pair using attention and grounding maps.

*Main Contributions and Findings:* Avoids reliance on coarse proposal sampling by computing dense alignment scores. Helps with **fine-grained temporal boundary detection**.

*Limitations and Future Directions:* Computationally expensive due to dense frame comparisons. Future work could employ **sparse or approximate matching strategies**.

*Relevance to Primary Paper:* Tackles **temporal precision and boundary errors**, as highlighted by the paper’s error categorization.

**63. VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval (Minuk Ma et al., 2020)**  
*Key Concepts and Taxonomy:* VLANet is a **weakly supervised model** that aligns videos and queries without explicit moment annotations. It leverages a **cross-modal attention mechanism** and ranking loss to identify the relevant video segments.

*Main Contributions and Findings:* It achieves robust performance by focusing on enhancing visual-semantic correspondence using a **multi-granularity attention strategy**. It surpasses earlier weakly-supervised models like TGA and WSLLN.

*Limitations and Future Directions:* As a weakly-supervised method, it still suffers in high ambiguity scenarios. Future extensions could use **contrastive learning or temporal consistency losses** to improve precision.

*Relevance to Primary Paper:* Addresses the **weakly supervised retrieval** frontier, which is a key challenge identified in the diagnostic study, particularly in reducing annotation cost while maintaining localization quality.

**64. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval (Mayu Otani et al., 2020)**  
*Key Concepts and Taxonomy:* This paper investigates the **hidden biases and evaluation inconsistencies** in the moment retrieval task, analyzing benchmark datasets and proposing more challenging protocols.

*Main Contributions and Findings:* It reveals that many benchmark test queries are **visually easy** or **textually biased**, which can inflate reported performance. The authors propose **hard query splits** to better assess real-world robustness.

*Limitations and Future Directions:* Primarily a diagnostic work, not proposing a new retrieval model. Future efforts could incorporate their harder splits into mainstream benchmarking.

*Relevance to Primary Paper:* Provides a critical lens similar to the primary paper's **error diagnosis goals**, reinforcing the need for fair and interpretable evaluation in moment retrieval.

**65. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips (Antoine Miech et al., 2019)**  
*Key Concepts and Taxonomy:* Introduces a **massive-scale pretraining dataset and method** for learning video-language embeddings from noisy narrations on instructional videos.

*Main Contributions and Findings:* Shows that weak supervision at scale can yield transferable embeddings that improve performance on downstream tasks like **moment retrieval, action recognition, and captioning**.

*Limitations and Future Directions:* Noisy narration-video alignment remains a challenge. Fine-tuning with domain-specific data is often necessary for optimal performance.

*Relevance to Primary Paper:* The pretrained embeddings from HowTo100M are used in many downstream temporal grounding models, enhancing **semantic alignment**, which the primary paper highlights as a major retrieval challenge.

**66. Weakly Supervised Video Moment Retrieval From Text Queries (Mithun et al., 2019)**  
*Key Concepts and Taxonomy:* This paper addresses the problem of localizing relevant moments in a video using natural language queries under weak supervision, meaning that the training data only provides video-level annotations rather than temporally aligned ones. It introduces a novel approach to combine textual and visual information using attention-based mechanisms to match video segments with query phrases.

*Main Contributions and Findings:* The paper proposes a two-stream network: one stream computes video features using visual cues, and the other aligns them with textual representations of the query. It incorporates self-attention and temporal attention to learn better temporal embeddings for weakly supervised training. Experimental results show significant improvements over baseline weakly supervised models.

*Limitations and Future Directions:* The reliance on weak supervision limits precision, especially for queries requiring fine-grained temporal alignment. Future work could explore better temporal modeling techniques or semi-supervised settings using limited segment-level annotations.

*Relevance to Primary Paper:* This work is critical in pushing the boundary of weakly supervised methods in video moment retrieval, offering techniques that inspire later fully-supervised and self-supervised approaches in the literature.

**67. Local-Global Video-Text Interactions for Temporal Grounding (Mun et al., 2020)**  
*Key Concepts and Taxonomy:*  
This work proposes a model that captures both **local and global interactions** between video and language for **temporal grounding** tasks. It introduces a **dual-level cross-modal interaction mechanism** to align fine-grained temporal video segments with natural language queries more effectively.

*Main Contributions and Findings:*

* Introduces a **local-global interaction module** combining local temporal features with global contextual features.
* Demonstrates the importance of hierarchical information for accurate localization.
* Achieves improved performance over state-of-the-art models on standard benchmarks like Charades-STA and ActivityNet Captions.

*Limitations and Future Directions:*

* Requires significant **computational overhead** due to multi-level processing.
* Struggles with very **long videos** where local features may still be ambiguous.
* Future work can explore more efficient attention-based architectures or dynamic query-conditioned temporal filters.

*Relevance to Primary Paper:*  
This paper reinforces the need for **fine-grained interaction modeling** in video-text grounding, contributing valuable architectural advancements used in subsequent hierarchical and attention-based models.

**68. Zero-shot Natural Language Video Localization (Nam et al., 2021)**  
*Key Concepts and Taxonomy:*  
This work addresses **Zero-shot Temporal Grounding**, where the system must localize query-relevant video segments without having seen any aligned training data from the same domain or query class. It proposes a **semantic space alignment framework** using pretrained models and domain adaptation.

*Main Contributions and Findings:*

* Utilizes **text and video embeddings** to construct a shared space without labeled video-query pairs.
* Incorporates **semantic relationships** from external language corpora.
* Shows that zero-shot grounding is feasible and yields promising results on standard benchmarks.

*Limitations and Future Directions:*

* Performance is lower than supervised methods.
* Relies on **semantic closeness** between seen and unseen concepts.
* Future work may involve more robust domain generalization and cross-domain adaptation techniques.

*Relevance to Primary Paper:*  
This work broadens the scope of temporal grounding by showing it can be **generalized to unseen classes** without supervision, pushing the task toward more real-world applications.

**69. Interventional Video Grounding with Dual Contrastive Learning (Nan et al., 2021)**  
*Key Concepts and Taxonomy:*  
Introduces a **causal learning perspective** to video grounding via an **interventional framework** using dual contrastive objectives. The method explicitly separates causal relevance from spurious correlations.

*Main Contributions and Findings:*

* Models query-video interaction with **interventional contrastive learning** to focus on truly relevant segments.
* Enhances grounding accuracy by **eliminating spurious correlations** through interventional data augmentation.

*Limitations and Future Directions:*

* Computational complexity due to intervention simulations.
* May not scale well to **longer or noisier videos**.
* Future extensions could involve real-world interventional data or semi-supervised setups.

*Relevance to Primary Paper:*  
This paper is cited as a **novel causal approach** that departs from purely correlation-based grounding methods, offering improved robustness.

**70. Interaction-Integrated Network for Natural Language Moment Localization (Ning et al., 2021)**  
*Key Concepts and Taxonomy:*  
Proposes an **Interaction-Integrated Network (I2N)** to enhance alignment between video and language through **multi-granularity interaction modeling**, leveraging both coarse and fine features.

*Main Contributions and Findings:*

* Combines **global video context** and **local temporal details** via hierarchical interaction blocks.
* Demonstrates state-of-the-art performance on datasets like Charades-STA and ActivityNet Captions.

*Limitations and Future Directions:*

* May suffer from **high memory consumption** due to multi-scale processing.
* Extension could include **lightweight variants** for mobile/edge deployment.

*Relevance to Primary Paper:*  
Supports the trend of **deep cross-modal interaction networks** for precise temporal grounding.

**71. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention (Rodriguez-Opazo et al., 2020)**  
*Key Concepts and Taxonomy:*  
This work eliminates the proposal-generation step by directly predicting the start and end moments using **guided attention mechanisms**.

*Main Contributions and Findings:*

* Proposes a **proposal-free framework**, reducing computation and latency.
* Uses a **guided attention encoder-decoder** that learns to focus on temporally relevant video regions.

*Limitations and Future Directions:*

* May not generalize well to **long or multi-action videos**.
* Incorporating external knowledge or memory modules could be a future step.

*Relevance to Primary Paper:*  
Demonstrates the feasibility of **lightweight, real-time** moment localization without relying on candidate proposals.

**72. Temporal Context Aggregation Network for Temporal Action Proposal Refinement (Qing et al., 2021)**  
*Key Concepts and Taxonomy:*  
Introduces the **Temporal Context Aggregation Network (TCANet)**, aimed at refining action proposals through multi-scale context aggregation.

*Main Contributions and Findings:*

* Uses **multi-scale temporal convolution** to capture broad contextual cues.
* Shows improvements in **temporal proposal precision** and recall.
* Achieves high mAP on benchmarks like THUMOS14.

*Limitations and Future Directions:*

* Mainly focused on **action detection**, not moment grounding from language.
* Future work could integrate **language-based grounding** into the same refinement framework.

*Relevance to Primary Paper:*  
Provides strong **proposal-level backbone enhancements**, useful for systems that rely on pre-generated temporal proposals.

**73. Video Object Grounding Using Semantic Roles in Language Description (Sadhu et al., 2020)**  
*Key Concepts and Taxonomy:*  
This paper introduces a novel method for grounding objects mentioned in video-related language descriptions by leveraging **semantic role labeling (SRL)** to identify relationships between objects and actions.

*Main Contributions and Findings:*

* Proposes **semantic-aware object grounding** that aligns verb-object pairs in text with visual regions.
* Demonstrates that modeling semantic roles improves precision for object-level grounding in videos.

*Limitations and Future Directions:*

* Relies on accurate SRL parsing, which may be noisy for complex queries.
* Future directions include **joint action-object grounding** and expanding to spatio-temporal localization.

*Relevance to Primary Paper:*  
Provides insights into fine-grained **object-level language grounding**, complementing moment-level video localization.

**74. What Actions Are Needed for Understanding Human Actions in Videos? (Sigurdsson et al., 2017)**  
*Key Concepts and Taxonomy:*  
This work focuses on the **compositional understanding of human actions**, studying what kinds of visual cues (e.g., objects, verbs, context) are essential for recognizing actions in video.

*Main Contributions and Findings:*

* Demonstrates that **object presence and interaction cues** significantly impact action understanding.
* Introduces an **action taxonomy** that decomposes actions into constituent parts.

*Limitations and Future Directions:*

* Operates mostly on **short, single-action videos**.
* Future work can expand to **temporal sequencing** and multi-action understanding.

*Relevance to Primary Paper:*  
Establishes a **semantic foundation** for why multimodal grounding is necessary—actions are not isolated but context-rich.

**75. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding (Sigurdsson et al., 2016)**  
*Key Concepts and Taxonomy:*  
Presents the **Charades dataset**, a large-scale collection of daily human activities captured in real homes using crowd workers.

*Main Contributions and Findings:*

* Innovates a **script-based data collection pipeline**.
* Provides diverse, realistic video data with **textual descriptions and labels** for action understanding.

*Limitations and Future Directions:*

* Home setting may not generalize to public or outdoor scenarios.
* Could be extended with **3D sensing or multi-modal capture**.

*Relevance to Primary Paper:*  
Charades is widely used as a benchmark for **video-language grounding** and moment retrieval tasks.

**76. VAL: Visual-Attention Action Localizer (Song and Han, 2018)**  
*Key Concepts and Taxonomy:*  
Introduces a **visual-attention-based localizer** for identifying temporal segments in videos corresponding to action descriptions.

*Main Contributions and Findings:*

* Proposes a novel **spatial-temporal attention mechanism**.
* Demonstrates improved action localization performance using attention alignment between video frames and textual action descriptions.

*Limitations and Future Directions:*

* Less emphasis on linguistic diversity in the queries.
* Future work may integrate **natural language understanding** modules for richer semantic mapping.

*Relevance to Primary Paper:*  
One of the early systems showing the value of **attention mechanisms** for aligning video segments and language.

**77. BSN++: Complementary Boundary Regressor with Scale-Balanced Relation Modeling for Temporal Action Proposal Generation (Su et al., 2021)**  
*Key Concepts and Taxonomy:*  
This paper presents **BSN++**, an enhanced proposal generator for temporal action detection, focusing on precise **segment boundary regression**.

*Main Contributions and Findings:*

* Introduces **scale-balanced modeling** to handle varying action durations.
* Achieves high performance on ActivityNet and THUMOS14 datasets.

*Limitations and Future Directions:*

* Focuses on **action proposals**, not moment localization from natural language.
* Future integration with **language-based models** could enable cross-modal grounding.

*Relevance to Primary Paper:*  
Provides **high-quality temporal proposals**, foundational for many video grounding pipelines.

**78. MABAN: Multi-Agent Boundary-Aware Network for Natural Language Moment Retrieval (Sun et al., 2021)**  
*Key Concepts and Taxonomy:*  
Introduces a **multi-agent collaborative framework** for natural language moment retrieval, where each agent specializes in refining start or end boundaries.

*Main Contributions and Findings:*

* Proposes **MABAN**, a model where agents interact and guide each other to improve temporal localization.
* Shows significant improvements over state-of-the-art on Charades-STA and TACoS datasets.

*Limitations and Future Directions:*

* Relies on well-separated action boundaries.
* Future improvements could include **uncertainty modeling** in ambiguous temporal contexts.

*Relevance to Primary Paper:*  
Pushes forward **boundary-aware localization**, complementing broader localization models with fine-grained boundary decisions.

**79. Reinforcement Learning: An Introduction (Sutton and Barto, 2018)**  
*Key Concepts and Taxonomy:*  
This is a foundational textbook introducing **reinforcement learning (RL)** concepts like Markov Decision Processes (MDPs), policy gradients, and Q-learning.

*Main Contributions and Findings:*

* Explains **value-based and policy-based RL methods** in detail.
* Forms the theoretical backbone for many RL-based approaches to video grounding.

*Limitations and Future Directions:*

* Mostly theoretical; lacks application-specific guidance.
* Future editions could include **video-language grounding applications** using RL.

*Relevance to Primary Paper:*  
Provides the **RL foundations** for models like **STRONG**, which apply spatio-temporal RL to video moment localization.

**80. Rethinking the Inception Architecture for Computer Vision (Szegedy et al., 2016)**  
*Key Concepts and Taxonomy:*  
Improves the **Inception architecture** by optimizing it for speed and accuracy in vision tasks.

*Main Contributions and Findings:*

* Introduces **Inception-v3**, with factorized convolutions and label-smoothing regularization.
* Achieves state-of-the-art performance on ImageNet.

*Limitations and Future Directions:*

* Originally designed for images, not videos.
* Extensions like **I3D** adapted it for video understanding.

*Relevance to Primary Paper:*  
Inception models serve as **visual encoders** in many video-language grounding systems, providing high-level frame features.

**81. *Logan: Latent Graph Co-attention Network for Weakly-Supervised Video Moment Retrieval (Tan et al., 2021)***  
*Key Concepts and Taxonomy:* This paper proposes the **Logan** framework for **weakly supervised video moment retrieval**, which avoids using precise moment annotations during training. It constructs **latent graphs** to model both video and language interactions and uses a **co-attention mechanism** to align them.

*Main Contributions and Findings:* The model innovatively represents inter- and intra-modal relationships via graph structures and integrates them using latent co-attention. This improves semantic alignment between the query and potential video moments, resulting in **state-of-the-art performance among weakly supervised methods**.

*Limitations and Future Directions:* Although it improves weakly supervised grounding, the model still lags behind supervised ones. Its performance heavily relies on the quality of video features extracted by pretrained encoders. Future work may explore integrating **end-to-end trainable visual encoders** and expanding to longer, more complex videos.

*Relevance to Primary Paper:* Logan is a key advancement in **weak supervision**, directly supporting the primary paper’s discussion on methods that reduce annotation costs while improving semantic alignment.

**82.** *Frame-wise Cross-modal Matching for Video Moment Retrieval (Tang et al., 2021)*  
**Key Concepts and Taxonomy:** This paper presents a **fine-grained frame-wise matching** approach for video moment retrieval, where individual video frames are aligned with query words to enhance temporal localization precision.  
**Main Contributions and Findings:** The method calculates **dense cross-modal similarity scores** between frames and words, aggregating them for moment prediction. It outperforms previous coarse-grained segment-level models, especially in localizing fine transitions in video content.  
**Limitations and Future Directions:** Frame-wise comparisons can be computationally expensive and sensitive to noisy frames or irrelevant query words. Future work could focus on **efficient attention pruning** or dynamic frame sampling.  
**Relevance to Primary Paper:** This work directly contributes to addressing **temporal misalignment** errors highlighted in the primary paper by refining frame-level semantic alignment strategies.

**83.** *Multi-Level Query Interaction for Temporal Language Grounding (Tang et al., 2021)*  
**Key Concepts and Taxonomy:** This model explores **multi-level query interaction**, where different linguistic abstractions (word, phrase, sentence) are aligned with temporal video segments to improve grounding.  
**Main Contributions and Findings:** The method captures **hierarchical language semantics** and integrates them via multi-scale attention across video segments. It improves localization for both short and long queries, showing adaptability across diverse grounding scenarios.  
**Limitations and Future Directions:** Handling complex linguistic hierarchies requires significant computation. The model may also struggle with **multi-event** queries without external memory. Future work can explore **language-conditioned memory modules**.  
**Relevance to Primary Paper:** Offers a nuanced take on **language modeling** in temporal grounding, directly responding to challenges around **query complexity** raised in the primary analysis.

**84.** *Learning Spatiotemporal Features With 3D Convolutional Networks (Tran et al., 2015)*  
**Key Concepts and Taxonomy:** This paper introduces **C3D**, a 3D convolutional neural network for learning **joint spatial and temporal features** from raw video frames. It is a foundational work in video representation learning.  
**Main Contributions and Findings:** C3D enables action and event modeling in videos by processing motion and appearance together. It generalizes well across video classification, detection, and retrieval tasks.  
**Limitations and Future Directions:** 3D CNNs are computationally intensive and struggle with long-range dependencies. Future improvements include integrating **temporal attention** or transformer-based extensions.  
**Relevance to Primary Paper:** C3D and its successors are widely used as **visual encoders** in video moment localization pipelines, forming the backbone of many models reviewed in the primary paper.

**85.** *Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition (Tu et al., 2019)*  
**Key Concepts and Taxonomy:** This work extends the VLAD aggregation method by emphasizing **action stages** in a video. It captures spatiotemporal dynamics by learning where and when actions happen.  
**Main Contributions and Findings:** By combining **stage-aware pooling** with temporal VLAD descriptors, it enhances action recognition performance on datasets like UCF101 and HMDB51.  
**Limitations and Future Directions:** Designed for recognition, not retrieval; adaptation to grounding tasks may require additional alignment components.  
**Relevance to Primary Paper:** Although not a grounding method, this paper informs **temporal feature aggregation** strategies for segment-level understanding in retrieval models.

**86.** *Dual Path Interaction Network for Video Moment Localization (Wang et al., 2020)*  
**Key Concepts and Taxonomy:** This paper proposes a **dual-path network** that models both **global-to-local** and **local-to-global interactions** between video segments and language queries.  
**Main Contributions and Findings:** The dual-path design ensures that both **contextual relevance** and **segment specificity** are modeled. It improves grounding performance over traditional single-path methods.  
**Limitations and Future Directions:** Requires careful fusion strategies to prevent information dilution. Future work may explore **adaptive fusion weights or confidence modeling**.  
**Relevance to Primary Paper:** Provides a strong architectural response to the diagnostic concern of **context neglect** in grounding models.

**87. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query (Wang et al., 2021)**  
*Key Concepts and Taxonomy:* This paper proposes a **multi-level interaction network** that hierarchically models query-to-video and video-to-query relations for more precise temporal localization.  
*Main Contributions and Findings:*

* It structures interactions at both **global and local levels**, enabling more accurate alignment across modalities.
* The hierarchical design improves grounding performance, especially in complex multi-action videos.  
  *Limitations and Future Directions:*
* Increased complexity can lead to **higher training costs** and possible overfitting on small datasets.
* Future work could explore more **lightweight architectures** with similar capabilities.  
  *Relevance to Primary Paper:* This model strengthens **cross-modal representation learning**, directly addressing concerns about alignment and context reasoning mentioned in the diagnostic study.

**88. Temporally Grounding Language Queries in Videos by Contextual Boundary-Aware Prediction (Wang et al., 2020)**  
*Key Concepts and Taxonomy:* This paper introduces a **context-aware boundary prediction** strategy for grounding queries in videos by explicitly modeling context around candidate segments.  
*Main Contributions and Findings:*

* Uses **semantic context modeling** to improve start/end boundary detection.
* Shows robust performance in temporal localization under both coarse and fine query conditions.  
  *Limitations and Future Directions:*
* Struggles with **long-term temporal dependencies** or queries referring to multiple scenes.
* Future work could include **memory-augmented architectures** to retain long-range context.  
  *Relevance to Primary Paper:* Directly addresses **boundary localization errors**, one of the key failure modes highlighted in the primary paper’s diagnostic framework.

**89. UntrimmedNets for Weakly Supervised Action Recognition and Detection (Wang et al., 2017)**  
*Key Concepts and Taxonomy:* Proposes **UntrimmedNets**, a framework for learning action detectors from untrimmed videos using weak video-level labels without precise temporal boundaries.  
*Main Contributions and Findings:*

* Combines **classification and selection modules** to detect the most relevant video segments.
* Pioneered weak supervision for action recognition and laid groundwork for weakly supervised moment retrieval.  
  *Limitations and Future Directions:*
* Focused on action categories, not natural language queries.
* Extensions needed for **semantic grounding** and richer query handling.  
  *Relevance to Primary Paper:* Serves as a precursor to **weakly-supervised grounding frameworks**, motivating models that avoid costly segment-level annotation.

**90. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model (Wang et al., 2019)**  
*Key Concepts and Taxonomy:* Introduces a **reinforcement learning**-based semantic matching model for grounding language queries in temporal video segments.  
*Main Contributions and Findings:*

* The agent interacts with video frames sequentially to locate relevant actions using **semantic similarity as a reward**.
* Demonstrates that RL can effectively model long-term query-video alignment.  
  *Limitations and Future Directions:*
* Sensitive to reward shaping and action-space design.
* Future work could integrate **hierarchical or multi-agent learning**.  
  *Relevance to Primary Paper:* Offers an early application of **RL in grounding**, which helps address issues of sequential misalignment noted in the diagnostic analysis.

**91. Weakly Supervised Temporal Adjacent Network for Language Grounding (Wang et al., 2021)**  
*Key Concepts and Taxonomy:* This paper introduces a **Temporal Adjacent Network** under weak supervision, which models the correlation between adjacent segments to better infer grounding positions.  
*Main Contributions and Findings:*

* Leverages segment adjacency to **propagate supervision signals** across time without fine annotations.
* Shows improved performance over traditional weakly supervised baselines.  
  *Limitations and Future Directions:*
* Assumes strong temporal correlation between adjacent moments, which might not hold in all scenarios.
* Could benefit from **adaptive adjacency modeling** based on content dynamics.  
  *Relevance to Primary Paper:* Directly contributes to **weakly supervised video grounding**, offering a structural solution to improve moment proposal quality without labeled timestamps.

**92. Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval (Zheng Wang et al., 2021)**  
*Key Concepts and Taxonomy:* This paper proposes a **co-occurrence alignment framework** for weakly supervised video moment retrieval. It focuses on the visual consistency between co-occurring concepts in video clips and query phrases.  
*Main Contributions and Findings:* The model learns to align visually co-occurring elements using a two-branch architecture—one for alignment learning and one for localization. It improves grounding performance without relying on segment-level annotations.  
*Limitations and Future Directions:* It assumes strong co-occurrence regularity, which may not hold in more diverse or long-form videos. Future work could explore more flexible alignment strategies or causal reasoning.  
*Relevance to Primary Paper:* Advances the direction of **weakly supervised semantic matching**, aligning closely with the primary paper’s emphasis on reducing annotation reliance.

**93. Multi-modal Circulant Fusion for Video-to-Language and Backward (Aming Wu and Yahong Han, 2018)**  
*Key Concepts and Taxonomy:* This work introduces a **circulant fusion layer** to integrate visual and language modalities in both directions—video-to-text and text-to-video.  
*Main Contributions and Findings:* The circulant fusion mechanism models bidirectional semantic relations effectively, enabling enhanced multimodal understanding and retrieval.  
*Limitations and Future Directions:* It is more focused on representation fusion than fine-grained temporal localization. Adaptation to moment-level tasks could be explored through frame-level attention modeling.  
*Relevance to Primary Paper:* The fusion strategy is referenced in several **cross-modal grounding models**, supporting robust multimodal alignment.

**94. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos (Jie Wu et al., 2020)**  
*Key Concepts and Taxonomy:* This paper frames grounding as a **Markov Decision Process**, where an agent learns to navigate temporal boundaries using reward signals derived from query-video alignment.  
*Main Contributions and Findings:* By modeling weak supervision as delayed rewards, the agent learns to localize moments without timestamp-level labels. Performance shows substantial improvements on ActivityNet and Charades-STA.  
*Limitations and Future Directions:* RL-based models are sensitive to **reward shaping and convergence stability**. Future approaches could explore curriculum-based rewards or hierarchical policies.  
*Relevance to Primary Paper:* Highlights the utility of **reinforcement learning** under weak supervision, reinforcing themes from the diagnostic study on training strategies.

**95. Tree-Structured Policy Based Progressive Reinforcement Learning for Temporally Language Grounding in Video (Jie Wu et al., 2020)**  
*Key Concepts and Taxonomy:* Builds on prior RL approaches by introducing a **tree-structured policy** that progressively refines temporal predictions in a coarse-to-fine manner.  
*Main Contributions and Findings:* This structured exploration allows the agent to focus on increasingly accurate intervals. It balances exploration and exploitation effectively, improving grounding precision.  
*Limitations and Future Directions:* The tree structure assumes predictable narrowing of the search space, which may not hold for complex, multi-action videos. Future improvements could include adaptive trees or attention-guided pruning.  
*Relevance to Primary Paper:* Reinforces the **progressive refinement strategies** discussed in the paper as effective tools for accurate video moment localization.

**96. Diving Into the Relations: Leveraging Semantic and Visual Structures For Video Moment Retrieval (Ziyue Wu et al., 2021)**  
*Key Concepts and Taxonomy:* This paper introduces a dual-branch network that leverages **semantic role structures** in language and **scene structures** in video for better grounding.  
*Main Contributions and Findings:* Through structured relation modeling, the network learns fine-grained matching between textual cues and visual contexts. This leads to improved retrieval performance across datasets.  
*Limitations and Future Directions:* Relation modeling increases complexity and training time. More efficient relation parsers or graph sparsification could be explored.  
*Relevance to Primary Paper:* Strongly ties into the **semantic misalignment** category, offering structural cues as a solution for more robust grounding.

**97. Natural Language Video Localization with Learnable Moment Proposals (Xiao et al., 2021)**  
*Key Concepts and Taxonomy:* This paper proposes a **two-stage localization model** where moment proposals are **learned end-to-end** rather than pre-defined. It introduces a learnable proposal module that generates temporal segments conditioned on the language query.  
*Main Contributions and Findings:* The method improves upon traditional proposal-based systems by jointly optimizing the proposal generation and moment alignment tasks. It significantly improves recall and precision on datasets like Charades-STA and ActivityNet.  
*Limitations and Future Directions:* Proposal generation quality depends heavily on query-video understanding, and might underperform with ambiguous or vague queries. Future work could incorporate **multi-scale or adaptive proposal mechanisms**.  
*Relevance to Primary Paper:* This model is directly relevant to addressing **proposal quality**, a key diagnostic dimension discussed in the primary paper.

**98. Boundary Proposal Network for Two-Stage Natural Language Video Localization (Xiao et al., 2021)**  
*Key Concepts and Taxonomy:* Builds on learnable proposals by introducing a **Boundary Proposal Network (BPN)** designed specifically for **precise boundary regression** in video moment retrieval.  
*Main Contributions and Findings:* BPN separates the localization process into two distinct stages—coarse proposal generation and boundary refinement. The boundary-aware module refines start and end times using context-enhanced features.  
*Limitations and Future Directions:* The two-stage structure adds computation and may still be sensitive to initial proposal errors. Future research may explore **joint boundary-prediction and alignment** frameworks.  
*Relevance to Primary Paper:* BPN directly tackles **temporal boundary mislocalization**, one of the main error modes in temporal grounding tasks.

**99. Dynamic Co-attention Networks for Question Answering (Xiong et al., 2017)**  
*Key Concepts and Taxonomy:* While focused on question answering, this paper introduces the **dynamic co-attention network**, a dual-attention mechanism where query and context attend to each other simultaneously.  
*Main Contributions and Findings:* The model effectively captures **bidirectional interactions** using co-attention and gating strategies, achieving strong results on QA benchmarks like SQuAD.  
*Limitations and Future Directions:* Originally designed for text QA, so it lacks temporal modeling or video alignment. Future adaptations could extend it to **multi-modal QA** or **video-language grounding**.  
*Relevance to Primary Paper:* Inspires many **co-attention-based architectures** in video grounding, especially those aligning temporal segments with queries.

**100. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection (Xu et al., 2017)**  
*Key Concepts and Taxonomy:* This work adapts **Faster R-CNN to the temporal domain**, proposing R-C3D, which performs **end-to-end action proposal and classification** in video using 3D convolutions.  
*Main Contributions and Findings:* It introduces a **temporal region proposal network** to generate candidate activity segments and classifies them jointly with a 3D ConvNet. It is highly efficient and accurate.  
*Limitations and Future Directions:* Originally designed for predefined action categories, not open-ended queries. Needs further extension for **language-based retrieval**.  
*Relevance to Primary Paper:* R-C3D is a backbone for several moment retrieval systems, offering strong **proposal generation baselines** critical for downstream grounding models.

**101. Multilevel Language and Vision Integration for Text-to-Clip Retrieval (Xu et al., 2019)**  
*Key Concepts and Taxonomy:* This paper introduces a multilevel integration framework that aligns textual and visual information at different levels of semantic abstraction for the task of text-to-clip retrieval.  
*Main Contributions and Findings:* By decomposing language and visual content hierarchically, the model captures both **coarse and fine-grained** associations. It improves retrieval accuracy on standard datasets by leveraging multiscale fusion mechanisms.  
*Limitations and Future Directions:* The approach assumes hierarchical alignment exists uniformly, which may not hold in multi-action scenes. Future work could benefit from **adaptive granularity modeling**.  
*Relevance to Primary Paper:* Provides a robust baseline for hierarchical language-vision interaction, directly addressing challenges in **cross-modal grounding precision** noted in the diagnostic study.

**102. Revisiting Anchor Mechanisms for Temporal Action Localization (Yang et al., 2020)**  
*Key Concepts and Taxonomy:* This paper re-evaluates traditional anchor-based methods in temporal action localization, proposing more adaptive and flexible anchor mechanisms with boundary refinement.  
*Main Contributions and Findings:* It introduces **optimized temporal anchors** and contextual boundary modules that significantly improve mAP on ActivityNet and THUMOS14 datasets.  
*Limitations and Future Directions:* The system is designed for **predefined action categories** and may not generalize well to **open-vocabulary or query-based** tasks. Future extensions could bridge anchor mechanisms with natural language grounding.  
*Relevance to Primary Paper:* Serves as a strong backbone for **proposal generation modules** used in text-driven temporal grounding models discussed in the primary paper.

**103. Local Correspondence Network for Weakly Supervised Temporal Sentence Grounding (Yang et al., 2021)**  
*Key Concepts and Taxonomy:* The paper proposes LCNet, which focuses on learning **local semantic correspondences** between video segments and sentence components under weak supervision.  
*Main Contributions and Findings:* Without relying on annotated timestamps, LCNet uses attention-based correspondence to align textual and visual cues effectively. It achieves competitive performance across multiple benchmarks.  
*Limitations and Future Directions:* Sensitive to the quality of segment features; performance can drop in videos with rapid transitions. Future improvements might include **temporal smoothing or global refinement modules**.  
*Relevance to Primary Paper:* LCNet exemplifies how **fine-grained weakly supervised alignment** can close the gap between fully supervised and weakly supervised models—core to the motivation of the primary paper.

# Links to Referenced paper and Sources

* **Localizing Moments in Video with Natural Language (Hendricks et al., 2017)**
  + **Paper:** [**https://openaccess.thecvf.com/content\_iccv\_2017/html/Hendricks\_Localizing\_Moments\_in\_ICCV\_2017\_paper.html**](https://openaccess.thecvf.com/content_iccv_2017/html/Hendricks_Localizing_Moments_in_ICCV_2017_paper.html)
* **TALL: Temporal Activity Localization via Language Query (Gao et al., 2017)**
  + **Paper:** [**https://openaccess.thecvf.com/content\_iccv\_2017/html/Gao\_TALL\_Temporal\_Activity\_ICCV\_2017\_paper.html**](https://openaccess.thecvf.com/content_iccv_2017/html/Gao_TALL_Temporal_Activity_ICCV_2017_paper.html)
* **Grounding Action Descriptions in Videos (Regneri et al., 2013)**
  + **Paper:** [**https://aclanthology.org/P13-1021/**](https://aclanthology.org/P13-1021/)
* **Dense-Captioning Events in Videos (Krishna et al., 2017)**
  + **arXiv:** [**https://arxiv.org/abs/1705.00754**](https://arxiv.org/abs/1705.00754)
* **Temporally Grounding Natural Sentence in Video (Chen et al., 2018)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3274237**](https://www.google.com/search?q=https://dl.acm.org/doi/10.1145/3274237&authuser=1)
* **Span-based Localizing Network for Natural Language Video Localization (Zhang et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/1909.02356**](https://arxiv.org/abs/1909.02356)
* **Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language (Zhang et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2003.06495**](https://arxiv.org/abs/2003.06495)
* **Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos (He et al., 2019)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3343031.3350921**](https://dl.acm.org/doi/10.1145/3343031.3350921)
* **Weakly Supervised Video Moment Retrieval From Text Queries (Mithun et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1903.04852**](https://arxiv.org/abs/1903.04852)
* **Learning Video Moment Retrieval Without a Single Annotated Video (Gao et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475676**](https://dl.acm.org/doi/10.1145/3474085.3475676)
* **Diagnosing Error in Temporal Action Detectors (Alwassel et al., 2018)**
  + **arXiv:** [**https://arxiv.org/abs/1806.01018**](https://arxiv.org/abs/1806.01018)
* **Continuum Robots for Medical Applications: A Survey (Burgner-Kahrs et al., 2015)**
  + **Journal:** [**https://ieeexplore.ieee.org/document/7101225**](https://www.google.com/search?q=https://ieeexplore.ieee.org/document/7101225&authuser=1)
* **STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization (Cao et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413727**](https://dl.acm.org/doi/10.1145/3394171.3413727)
* **Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization (Cao et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413729**](https://dl.acm.org/doi/10.1145/3394171.3413729)
* **Localizing Natural Language in Videos (Chen et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1809.03262**](https://arxiv.org/abs/1809.03262)
* **Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos (Shaoxiang Chen et al., 2020)**
  + **Paper:** [**https://openaccess.thecvf.com/content\_WACV\_2020/html/Chen\_Learning\_Modality\_Interaction\_for\_Temporal\_Sentence\_Localization\_and\_Event\_Captioning\_WACV\_2020\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_WACV_2020/html/Chen_Learning_Modality_Interaction_for_Temporal_Sentence_Localization_and_Event_Captioning_WACV_2020_paper.html&authuser=1)
* **Semantic Proposal for Activity Localization in Videos via Sentence Query (Shaoxiang Chen & Yugang Jiang, 2019)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3343031.3350953**](https://dl.acm.org/doi/10.1145/3343031.3350953)
* **End-to-End Multi-Modal Video Temporal Grounding (Yi-Wen Chen et al., 2021)**
  + **Paper:** [**https://openaccess.thecvf.com/content/WACV2021/html/Chen\_End-to-End\_Multi-Modal\_Video\_Temporal\_Grounding\_WACV\_2021\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content/WACV2021/html/Chen_End-to-End_Multi-Modal_Video_Temporal_Grounding_WACV_2021_paper.html&authuser=1)
* **Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video (Zhenfang Chen et al., 2019)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3343031.3350942**](https://dl.acm.org/doi/10.1145/3343031.3350942)
* **A Review of Motion Planning for Highway Autonomous Driving (Claussmann et al., 2020)**
  + **Journal:** [**https://ieeexplore.ieee.org/document/8823338**](https://ieeexplore.ieee.org/document/8823338)
* **Temporal Localization of Moments in Video Collections with Natural Language (Escorcia et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1904.05849**](https://arxiv.org/abs/1904.05849)
* **Relation-aware Video Reading Comprehension for Temporal Language Grounding (Jialin Gao et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475624**](https://dl.acm.org/doi/10.1145/3474085.3475624)
* **Fast Video Moment Retrieval (Junyu Gao and Changsheng Xu, 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475351**](https://dl.acm.org/doi/10.1145/3474085.3475351)
* **WSLLN: Weakly Supervised Natural Language Localization Networks (Mingfei Gao et al., 2019)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/8953930/**](https://ieeexplore.ieee.org/document/8953930/)
* **MAC: Mining Activity Concepts for Language-Based Temporal Localization (Runzhou Ge et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1904.05652**](https://arxiv.org/abs/1904.05652)
* **ExCL: Extractive Clip Localization Using Natural Language Descriptions (Soham Ghosh et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1904.03334**](https://arxiv.org/abs/1904.03334)
* **Tripping Through Time: Efficient Localization of Activities in Videos (Meera Hahn et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/1911.08223**](https://arxiv.org/abs/1911.08223)
* **Localizing Moments in Video with Temporal Language (Lisa Anne Hendricks et al., 2018)**
  + **arXiv:** [**https://arxiv.org/abs/1708.01641**](https://arxiv.org/abs/1708.01641)
* **Natural Language Object Retrieval (Ronghang Hu et al., 2016)**
  + **arXiv:** [**https://arxiv.org/abs/1603.04189**](https://arxiv.org/abs/1603.04189)
* **Video Moment Localization via Deep Cross-Modal Hashing (Yupeng Hu et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9428540/**](https://ieeexplore.ieee.org/document/9428540/)
* **Coarse-to-Fine Semantic Alignment for Cross-Modal Moment Localization (Yupeng Hu et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9368564/**](https://www.google.com/search?q=https://ieeexplore.ieee.org/document/9368564/&authuser=1)
* **Cross-Modal Video Moment Retrieval with Spatial and Language-Temporal Attention (Bin Jiang et al., 2019)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3323873.3325026**](https://dl.acm.org/doi/10.1145/3323873.3325026)
* **Skip-Thought Vectors (Ryan Kiros et al., 2015)**
  + **arXiv:** [**https://arxiv.org/abs/1506.06726**](https://arxiv.org/abs/1506.06726)
* **Bidirectional Single-Stream Temporal Sentence Query Localization in Untrimmed Videos (Cheng Li et al., 2019)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3343031.3351047**](https://dl.acm.org/doi/10.1145/3343031.3351047)
* **Multi-scale 2D Representation Learning for Weakly-supervised Moment Retrieval (Ding Li et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475625**](https://dl.acm.org/doi/10.1145/3474085.3475625)
* **Global-local Temporal Representations for Video Person Re-identification (Jianing Li et al., 2019)**
  + **Paper:** [**https://openaccess.thecvf.com/content\_ICCV\_2019/html/Li\_Global-Local\_Temporal\_Representations\_for\_Video\_Person\_Re-Identification\_ICCV\_2019\_paper.html**](https://openaccess.thecvf.com/content_ICCV_2019/html/Li_Global-Local_Temporal_Representations_for_Video_Person_Re-Identification_ICCV_2019_paper.html)
* **Multi-scale Temporal Cues Learning for Video Person Reidentification (Jianing Li et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413571**](https://dl.acm.org/doi/10.1145/3394171.3413571)
* **43. Proposal-Free Video Grounding with Contextual Pyramid Network (Kun Li et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475560**](https://dl.acm.org/doi/10.1145/3474085.3475560)
* **44. Local-enhanced Interaction for Temporal Moment Localization (Guoqiang Liang et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9666013/**](https://ieeexplore.ieee.org/document/9666013/)
* **45. Continuous Control with Deep Reinforcement Learning (Lillicrap et al., 2016)**
  + **arXiv:** [**https://arxiv.org/abs/1509.02971**](https://arxiv.org/abs/1509.02971)
* **46. Single Shot Temporal Action Detection (Tianwei Lin et al., 2017)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3123266.3123305**](https://www.google.com/search?q=https://dl.acm.org/doi/10.1145/3123266.3123305&authuser=1)
* **47. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network (Zhijie Lin et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413725**](https://dl.acm.org/doi/10.1145/3394171.3413725)
* **48. Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction (Z. Lin et al., 2020)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9156715/**](https://ieeexplore.ieee.org/document/9156715/)
* **49. Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos (Bingbin Liu et al., 2018)**
  + **arXiv:** [**https://arxiv.org/abs/1811.07123**](https://arxiv.org/abs/1811.07123)
* **50. Reasoning Step-by-Step: Temporal Sentence Localization in Videos via Deep Rectification-Modulation Network (Daizong Liu et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2004.14818**](https://arxiv.org/abs/2004.14818)
* **51. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos (Daizong Liu et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9368566/**](https://www.google.com/search?q=https://ieeexplore.ieee.org/document/9368566/&authuser=1)
* **52. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding (Daizong Liu et al., 2021)**
  + **arXiv:** [**https://arxiv.org/abs/2108.05849**](https://arxiv.org/abs/2108.05849)
* **53. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization (Daizong Liu et al., 2020)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9156716/**](https://ieeexplore.ieee.org/document/9156716/)
* **54. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding (Daizong Liu et al., 2021)**
  + **arXiv:** [**https://arxiv.org/abs/2103.09241**](https://arxiv.org/abs/2103.09241)
* **55. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning (M. Liu et al., 2019)**
  + **Journal:** [**https://ieeexplore.ieee.org/document/8693951/**](https://www.google.com/search?q=https://ieeexplore.ieee.org/document/8693951/&authuser=1)
* **56. Iterative Local-Global Collaboration Learning Towards One-Shot Video Person Re-Identification (M. Liu et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413570**](https://dl.acm.org/doi/10.1145/3394171.3413570)
* **57. Attentive Moment Retrieval in Videos (Meng Liu et al., 2018)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3209978.3210042**](https://dl.acm.org/doi/10.1145/3209978.3210042)
* **58. Cross-Modal Moment Localization in Videos (Meng Liu et al., 2018)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3240508.3240590**](https://dl.acm.org/doi/10.1145/3240508.3240590)
* **59. SSD: Single Shot MultiBox Detector (Wei Liu et al., 2016)**
  + **arXiv:** [**https://arxiv.org/abs/1512.02325**](https://arxiv.org/abs/1512.02325)
* **60. A Survey on Natural Language Video Localization (Xinfang Liu et al., 2021)**
  + **arXiv:** [**https://arxiv.org/abs/2101.01214**](https://arxiv.org/abs/2101.01214)
* **61. Single-shot Semantic Matching Network for Moment Localization in Videos (Xinfang Liu et al., 2021)**
  + **Journal:** [**https://www.sciencedirect.com/science/article/abs/pii/S092523122100067X**](https://www.google.com/search?q=https://www.sciencedirect.com/science/article/abs/pii/S092523122100067X&authuser=1)
* **62. Debug: Dense Bottom-Up Grounding Approach for Natural Language Video Localization (Chujie Lu et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1908.00488**](https://arxiv.org/abs/1908.00488)
* **63. VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval (Minuk Ma et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413724**](https://dl.acm.org/doi/10.1145/3394171.3413724)
* **64. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval (Mayu Otani et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2004.03372**](https://arxiv.org/abs/2004.03372)
* **65. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips (Antoine Miech et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1906.03327**](https://arxiv.org/abs/1906.03327)
* **66. Weakly Supervised Video Moment Retrieval From Text Queries (Mithun et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1903.04852**](https://arxiv.org/abs/1903.04852)
* **67. Local-Global Video-Text Interactions for Temporal Grounding (Mun et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/1911.09524**](https://arxiv.org/abs/1911.09524)
* **68. Zero-shot Natural Language Video Localization (Nam et al., 2021)**
  + **arXiv:** [**https://arxiv.org/abs/2103.08473**](https://www.google.com/search?q=https://arxiv.org/abs/2103.08473&authuser=1)
* **69. Interventional Video Grounding with Dual Contrastive Learning (Nan et al., 2021)**
  + **arXiv:** [**https://arxiv.org/abs/2108.06734**](https://arxiv.org/abs/2108.06734)
* **70. Interaction-Integrated Network for Natural Language Moment Localization (Ning et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475626**](https://dl.acm.org/doi/10.1145/3474085.3475626)
* **71. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention (Rodriguez-Opazo et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2003.05153**](https://arxiv.org/abs/2003.05153)
* **72. Temporal Context Aggregation Network for Temporal Action Proposal Refinement (Qing et al., 2021)**
  + **arXiv:** [**https://arxiv.org/abs/2103.09244**](https://arxiv.org/abs/2103.09244)
* **73. Video Object Grounding Using Semantic Roles in Language Description (Sadhu et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/1912.01278**](https://arxiv.org/abs/1912.01278)
* **74. What Actions Are Needed for Understanding Human Actions in Videos? (Sigurdsson et al., 2017)**
  + **arXiv:** [**https://arxiv.org/abs/1705.02665**](https://arxiv.org/abs/1705.02665)
* **75. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding (Sigurdsson et al., 2016)**
  + **arXiv:** [**https://arxiv.org/abs/1604.01753**](https://arxiv.org/abs/1604.01753)
* **76. VAL: Visual-Attention Action Localizer (Song and Han, 2018)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/8578912/**](https://ieeexplore.ieee.org/document/8578912/)
* **77. BSN++: Complementary Boundary Regressor with Scale-Balanced Relation Modeling for Temporal Action Proposal Generation (Su et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475559**](https://dl.acm.org/doi/10.1145/3474085.3475559)
* **78. MABAN: Multi-Agent Boundary-Aware Network for Natural Language Moment Retrieval (Sun et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475561**](https://dl.acm.org/doi/10.1145/3474085.3475561)
* **79. Reinforcement Learning: An Introduction (Sutton and Barto, 2018)**
  + **Book:** [**http://incompleteideas.net/book/the-book-2nd.html**](http://incompleteideas.net/book/the-book-2nd.html)
* **80. Rethinking the Inception Architecture for Computer Vision (Szegedy et al., 2016)**
  + **arXiv:** [**https://arxiv.org/abs/1512.00567**](https://arxiv.org/abs/1512.00567)
* **81. Logan: Latent Graph Co-attention Network for Weakly-Supervised Video Moment Retrieval (Tan et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475562**](https://dl.acm.org/doi/10.1145/3474085.3475562)
* **82. Frame-wise Cross-modal Matching for Video Moment Retrieval (Tang et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475627**](https://dl.acm.org/doi/10.1145/3474085.3475627)
* **83. Multi-Level Query Interaction for Temporal Language Grounding (Tang et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9665621/**](https://ieeexplore.ieee.org/document/9665621/)
* **84. Learning Spatiotemporal Features With 3D Convolutional Networks (Tran et al., 2015)**
  + **arXiv:** [**https://arxiv.org/abs/1412.0767**](https://arxiv.org/abs/1412.0767)
* **85. Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition (Tu et al., 2019)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3343031.3350920**](https://dl.acm.org/doi/10.1145/3343031.3350920)
* **86. Dual Path Interaction Network for Video Moment Localization (Wang et al., 2020)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3394171.3413701**](https://dl.acm.org/doi/10.1145/3394171.3413701)
* **87. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query (Wang et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475563**](https://dl.acm.org/doi/10.1145/3474085.3475563)
* **88. Temporally Grounding Language Queries in Videos by Contextual Boundary-Aware Prediction (Wang et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2002.11326**](https://arxiv.org/abs/2002.11326)
* **89. UntrimmedNets for Weakly Supervised Action Recognition and Detection (Wang et al., 2017)**
  + **arXiv:** [**https://arxiv.org/abs/1701.01132**](https://arxiv.org/abs/1701.01132)
* **90. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model (Wang et al., 2019)**
  + **Paper:** [**https://openaccess.thecvf.com/content\_CVPR\_2019/html/Wang\_Language-Driven\_Temporal\_Activity\_Localization\_A\_Semantic\_Matching\_Reinforcement\_Learning\_Model\_CVPR\_2019\_paper.html**](https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Language-Driven_Temporal_Activity_Localization_A_Semantic_Matching_Reinforcement_Learning_Model_CVPR_2019_paper.html)
* **91. Weakly Supervised Temporal Adjacent Network for Language Grounding (Wang et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9665622/**](https://ieeexplore.ieee.org/document/9665622/)
* **92. Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval (Zheng Wang et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475564**](https://dl.acm.org/doi/10.1145/3474085.3475564)
* **93. Multi-modal Circulant Fusion for Video-to-Language and Backward (Aming Wu and Yahong Han, 2018)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3240508.3240591**](https://dl.acm.org/doi/10.1145/3240508.3240591)
* **94. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos (Jie Wu et al., 2020)**
  + **Paper:** [**https://openaccess.thecvf.com/content\_CVPR\_2020/html/Wu\_Reinforcement\_Learning\_for\_Weakly\_Supervised\_Temporal\_Grounding\_of\_Natural\_Language\_CVPR\_2020\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2020/html/Wu_Reinforcement_Learning_for_Weakly_Supervised_Temporal_Grounding_of_Natural_Language_CVPR_2020_paper.html&authuser=1)
* **95. Tree-Structured Policy Based Progressive Reinforcement Learning for Temporally Language Grounding in Video (Jie Wu et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2002.04351**](https://arxiv.org/abs/2002.04351)
* **96. Diving Into the Relations: Leveraging Semantic and Visual Structures For Video Moment Retrieval (Ziyue Wu et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475565**](https://dl.acm.org/doi/10.1145/3474085.3475565)
* **97. Natural Language Video Localization with Learnable Moment Proposals (Xiao et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9665623/**](https://ieeexplore.ieee.org/document/9665623/)
* **98. Boundary Proposal Network for Two-Stage Natural Language Video Localization (Xiao et al., 2021)**
  + **Paper:** [**https://ieeexplore.ieee.org/document/9665624/**](https://ieeexplore.ieee.org/document/9665624/)
* **99. Dynamic Co-attention Networks for Question Answering (Xiong et al., 2017)**
  + **arXiv:** [**https://arxiv.org/abs/1611.01604**](https://arxiv.org/abs/1611.01604)
* **100. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection (Xu et al., 2017)**
  + **arXiv:** [**https://arxiv.org/abs/1703.07814**](https://arxiv.org/abs/1703.07814)
* **101. Multilevel Language and Vision Integration for Text-to-Clip Retrieval (Xu et al., 2019)**
  + **arXiv:** [**https://arxiv.org/abs/1904.08518**](https://arxiv.org/abs/1904.08518)
* **102. Revisiting Anchor Mechanisms for Temporal Action Localization (Yang et al., 2020)**
  + **arXiv:** [**https://arxiv.org/abs/2009.02517**](https://arxiv.org/abs/2009.02517)
* **103. Local Correspondence Network for Weakly Supervised Temporal Sentence Grounding (Yang et al., 2021)**
  + **Paper:** [**https://dl.acm.org/doi/10.1145/3474085.3475628**](https://dl.acm.org/doi/10.1145/3474085.3475628)