Summaries of 'A Survey on Natural Language Video Localization' and its References

# https://arxiv.org/pdf/2104.00234

# Primary Paper: A Survey on Natural Language Video Localization (Liu et al., 2021)

## Key Concepts and Methodology

This survey provides a comprehensive and structured review of Natural Language Video Localization (NLVL), a task that aims to identify a specific temporal segment in a video that corresponds to a given natural language query. The paper establishes a canonical three-stage pipeline that characterizes most NLVL models: (1) Feature Extraction, where raw video and text are converted into numerical representations; (2) Cross-Modal Interaction, where features from both modalities are fused and aligned to learn joint representations; and (3) Localization, where the final temporal boundaries of the target segment are predicted. The core taxonomy of the field is presented, dividing methods into supervised and weakly-supervised learning. Supervised methods, which rely on precise start and end time annotations for training, are further broken down into five distinct sub-categories based on their localization strategy: Proposal-Based, Dense Locators, Single-Shot, Reinforcement Learning, and Boundary-Aware methods. This classification provides a clear framework for understanding the evolution and architectural diversity of NLVL research.

## Main Contributions and Findings

The primary contribution of this work is its meticulous and fine-grained organization of the NLVL landscape. It synthesizes a large body of literature into a coherent taxonomy, making the field more accessible to newcomers. The paper provides a thorough analysis of the strengths, weaknesses, and underlying assumptions of various approaches. It also offers a consolidated performance comparison of state-of-the-art models across four standard benchmarks: DiDeMo, TACOS, Charades-STA, and ActivityNet Captions. A key finding is that while significant progress has been made, the accuracy of current models is often insufficient for reliable real-world applications, highlighting a gap between academic research and practical deployment.

## Limitations and Broader Context

The survey identifies several critical challenges and limitations that hinder progress in the field. A major issue is the dataset bias; existing benchmarks are dominated by relatively short videos (a few minutes at most), which does not reflect the scale of real-world videos that can be hours long, posing a significant robustness and scalability problem. Another key limitation is the lack of model interpretability, as most deep learning models operate as 'black boxes,' making it difficult to understand their decision-making process. The paper also points out that many NLVL methods rely on outdated visual and textual feature extractors (e.g., VGG, C3D, GloVe), potentially creating a performance bottleneck and preventing the field from benefiting from more recent advances in representation learning, such as Transformer-based models.

## Relevance to Primary Survey

This is the central survey paper that provides the conceptual framework, taxonomy, and context for all other referenced works. It defines the problem space and organizes the various research contributions into a structured narrative.

# Reference [1]: Temporal Action Localization in Untrimmed Videos via Multi-Stage CNNs (Shou et al., 2016)

## Key Concepts and Methodology

This paper introduces a foundational framework for Temporal Action Localization using a Multi-Stage Convolutional Neural Network (CNN). The proposed method operates by first generating candidate temporal segments that might contain an action. These segments are then passed through a classification CNN to identify the action category. A subsequent localization CNN then refines the start and end boundaries of these proposals. This decoupling of proposal, classification, and localization into distinct stages allows the network to effectively handle actions of varying lengths and complexities within untrimmed videos.

## Relevance to Primary Survey

Cited as a classic example of Temporal Action Localization to draw a clear distinction with NLVL. While this work localizes actions from a predefined set of categories (e.g., 'high jump', 'diving'), NLVL addresses open-ended, descriptive language queries.

# Reference [2]: Cascaded Boundary Regression for Temporal Action Detection (Gao et al., 2017)

## Key Concepts and Methodology

This research enhances Temporal Action Localization by proposing a Cascaded Boundary Regression model. The core idea is to iteratively refine the temporal boundaries of action proposals. An initial set of proposals is generated, and a network then predicts adjustments to their start and end times. This process is repeated in a cascaded manner, with each stage taking the refined boundaries from the previous one as input, leading to progressively more accurate localization.

## Relevance to Primary Survey

Referenced as another key paper in the Temporal Action Localization domain. The survey points out that the architectural ideas from this related field, such as boundary regression, have been influential and often adapted by NLVL models.

# Reference [3]: Rethinking the Faster R-CNN Architecture for Temporal Action Localization (Chao et al., 2018)

## Key Concepts and Methodology

This paper adapts the highly successful Faster R-CNN object detection framework for the task of Temporal Action Localization. It re-purposes the core components of Faster R-CNN for the temporal domain: a Region Proposal Network (RPN) is used to propose candidate temporal segments (actions) instead of spatial regions (objects), and a detection head classifies these temporal proposals and refines their boundaries.

## Relevance to Primary Survey

Used as a reference to define the related field of Temporal Action Localization. This work is a prime example of how concepts from object detection (a spatial task) can inspire architectures for temporal tasks, a theme also present in NLVL.

# Reference [4]: End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos (Buch et al., 2019)

## Key Concepts and Methodology

This research proposes a unified, end-to-end model for Temporal Action Localization that diverges from multi-stage pipelines. It uses a single-stream 3D convolutional network that simultaneously predicts action proposals and classifies them. By integrating these steps into one network, the model can be trained jointly, potentially learning more effective representations for both tasks.

## Relevance to Primary Survey

Cited to help differentiate NLVL from the more traditional task of Temporal Action Localization. The end-to-end philosophy in this paper contrasts with many early NLVL models that used multi-stage, proposal-based approaches.

# Reference [5]: Gaussian Temporal Awareness Networks for Action Localization (Long et al., 2019)

## Key Concepts and Methodology

This paper introduces a novel approach to Temporal Action Localization by modeling the temporal structure of actions using Gaussian distributions. The network is designed to predict the center of an action instance as well as its temporal duration (standard deviation), effectively representing each action as a temporal Gaussian bell curve. This provides a soft, probabilistic approach to localization rather than predicting hard boundaries.

## Relevance to Primary Survey

Listed as a reference to provide a clear distinction between NLVL and Temporal Action Localization, showcasing the diverse modeling techniques within the latter field.

# Reference [6]: Visual to Text: Survey of Image and Video Captioning (Li et al., 2019)

## Key Concepts and Methodology

This paper provides a comprehensive survey of the field of image and video captioning. The primary goal of this field is to automatically generate a natural language description (a caption) that accurately describes the content of an image or video. The survey covers encoder-decoder architectures, the role of attention mechanisms, and various evaluation metrics used to assess the quality of generated captions.

## Relevance to Primary Survey

Cited to define Video Captioning as a related but inverse task to NLVL. While NLVL goes from text to video (localization), captioning goes from video to text (description). Despite their opposing goals, they share deep technical similarities in multimodal data processing and alignment.

# Reference [7]: Reconstruction Network for Video Captioning (Wang et al., 2018)

## Key Concepts and Methodology

This work in Video Captioning introduces a reconstruction network to improve the quality of generated captions. After generating a caption from a video, a reconstructor module attempts to recreate the original video features from the generated caption. A successful reconstruction implies that the caption has captured the essential semantic information of the video, acting as a form of self-supervision.

## Relevance to Primary Survey

Referenced to illustrate the field of Video Captioning. The survey notes that this idea of a cyclical relationship (video-to-text and text-to-video) has inspired dual-task models in NLVL that perform both localization and captioning simultaneously.

# Reference [8]: Video Captioning with Transferred Semantic Attributes (Pan et al., 2017)

## Key Concepts and Methodology

This contribution to the Video Captioning field focuses on leveraging high-level semantic attributes to generate more descriptive captions. The model is first trained to predict a set of semantic attributes (e.g., 'running', 'outdoors') from the video. These attributes are then fed into the language model to guide the caption generation process, resulting in richer and more accurate descriptions.

## Relevance to Primary Survey

Used as a reference to define Video Captioning and highlight the shared challenge of cross-modal alignment. The idea of using intermediate semantic concepts to bridge the modality gap is also explored in some NLVL models.

# Reference [9]: Video Captioning with Attention-Based LSTM and Semantic Consistency (Gao et al., 2017)

## Key Concepts and Methodology

This paper presents a video captioning model that utilizes an attention-based Long Short-Term Memory (LSTM) network. The attention mechanism allows the model to dynamically focus on the most relevant spatial regions of the video frames as it generates each word in the caption. The model also incorporates a semantic consistency objective to ensure the generated captions are globally coherent.

## Relevance to Primary Survey

Cited as an example of Video Captioning research. The survey explicitly notes that many NLVL models borrow or adapt network structures, particularly attention mechanisms, from the video captioning domain.

# Reference [10]: Video Captioning via Hierarchical Reinforcement Learning (Wang et al., 2018)

## Key Concepts and Methodology

This work applies a hierarchical reinforcement learning (HRL) framework to the task of video captioning. The HRL agent has a high-level manager policy that sets sub-goals and a low-level worker policy that generates language to satisfy those goals. This structure helps in generating longer, more coherent, and paragraph-like descriptions for videos, addressing the challenge of long-term temporal dependencies.

## Relevance to Primary Survey

Cited as an example of reinforcement learning's successful application in the broader field of video understanding. This success provides a strong motivation for exploring RL-based approaches for the NLVL task as well.

# Reference [11]: Weakly Supervised Dense Video Captioning (Shen et al., 2017)

## Key Concepts and Methodology

This research addresses dense video captioning (generating descriptions for multiple events in a video) under a weakly supervised setting. The model is only given a set of sentences that describe events in the video but not their temporal locations. It learns to both identify the event segments and assign the correct descriptions to them simultaneously.

## Relevance to Primary Survey

Cited as a Video Captioning paper to highlight the similarities in techniques shared with NLVL. The weakly-supervised setting here is analogous to the weakly-supervised NLVL task, where models must align sentences and video segments without explicit temporal annotations.

# Reference [12]: SST: Single-Stream Temporal Action Proposals (Buch et al., 2017)

## Key Concepts and Methodology

This paper focuses on the task of Temporal Action Proposals, which aims to generate a set of high-quality temporal segments from a video that are likely to contain an action. It introduces a single-stream temporal action proposal network that is computationally efficient and effective at identifying semantically meaningful action segments without classifying them.

## Relevance to Primary Survey

Referenced to define the Temporal Action Proposals task. The survey notes that a significant portion of early, proposal-based NLVL research directly borrowed or built upon methods like this to generate the initial candidate segments for localization.

# Reference [13]: TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals (Gao et al., 2017)

## Key Concepts and Methodology

This work proposes a Temporal Unit Regression Network (TURN) for generating high-quality temporal action proposals. It processes video features at multiple temporal scales and uses a regression module to predict precise boundaries for potential actions. A key contribution was also the establishment of standardized training and testing splits for the TACOS and Charades-STA datasets, which became crucial for fair comparison in subsequent research.

## Relevance to Primary Survey

Cited as an influential work in Temporal Action Proposals whose methods are often leveraged in NLVL. Its contribution of standardizing dataset splits was critical for the development and evaluation of many NLVL models.

# Reference [14]: DAPS: Deep Action Proposals for Action Understanding (Escorcia et al., 2016)

## Key Concepts and Methodology

A paper in the Temporal Action Proposals field that introduces a deep learning model to generate action proposals. The method evaluates video sub-sequences of various lengths and assigns an 'actionness' score to each, indicating the likelihood that it contains an action. This allows the model to identify semantically relevant segments in a class-agnostic manner.

## Relevance to Primary Survey

Cited as another example of a Temporal Action Proposal method that has been widely adopted or used as a baseline by researchers in the NLVL community.

# Reference [15]: Play and Rewind: Context-Aware Video Temporal Action Proposals (Gao et al., 2020)

## Key Concepts and Methodology

This paper enhances Temporal Action Proposals by introducing a context-aware model. The 'Play and Rewind' mechanism processes the video in both forward and backward directions, allowing the model to capture richer temporal context. This helps in distinguishing ambiguous actions and generating more accurate proposals, especially for activities that are defined by the events that precede or follow them.

## Relevance to Primary Survey

Used as a reference to define the Temporal Action Proposals task and its strong connection to NLVL. Context-aware proposals are particularly relevant for localizing complex language queries that describe sequential actions.

# Reference [16]: Learning Spatiotemporal Features with 3D Convolutional Networks (Tran et al., 2015)

## Key Concepts and Methodology

This seminal paper introduced the C3D (Convolutional 3D) network, a 3D convolutional architecture for video feature extraction. Unlike 2D CNNs that process static images, C3D uses 3D kernels that operate over both spatial and temporal dimensions (height, width, and time). This allows the network to directly learn motion and appearance features from short video clips.

## Main Contributions and Findings

C3D established a powerful and straightforward method for learning spatiotemporal features from raw video. It was pretrained on a large-scale sports video dataset and became a de facto standard for video representation in a wide range of video understanding tasks.

## Relevance to Primary Survey

Cited as one of the most common and foundational clip-level feature extractors used by NLVL models. The survey notes that many pioneering NLVL models, such as TALL and ACL, rely on pre-trained C3D features to represent the visual content of the video.

# Reference [17]: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (Carreira & Zisserman, 2017)

## Key Concepts and Methodology

This paper introduced the I3D (Inflated 3D ConvNet) model for action recognition. The core idea is to take a successful 2D image classification network (like Inception-v1) and 'inflate' its 2D kernels (e.g., k x k) into 3D kernels (e.g., t x k x k) by repeating them across the time dimension. This allows for leveraging powerful pre-trained image models for video tasks.

## Main Contributions and Findings

I3D, along with the large-scale Kinetics dataset introduced in the same paper, set a new state-of-the-art for action recognition and became a highly influential video feature extractor. It demonstrated the effectiveness of transfer learning from image to video domains.

## Relevance to Primary Survey

Cited as a primary clip-level feature extractor used in NLVL, often seen as a more powerful alternative to C3D. The survey mentions that modern NLVL models like DORi use I3D features to capture more robust dynamic information from video.

# Reference [18]: Very Deep Convolutional Networks for Large-Scale Image Recognition (Simonyan & Zisserman, 2014)

## Key Concepts and Methodology

This landmark paper introduced the VGG network, a very deep 2D convolutional neural network for image recognition. The key insight of VGG was to show that a very deep network (16-19 layers) composed of simple 3x3 convolutional filters could achieve excellent performance. This uniform and deep architecture made it easy to scale and adapt.

## Main Contributions and Findings

VGG demonstrated that network depth is a critical factor for performance in computer vision. It became a standard backbone architecture for a multitude of vision tasks, far beyond simple image classification.

## Relevance to Primary Survey

Cited as a common frame-level feature extractor used in NLVL. By applying a pre-trained VGG network to individual frames of a video, models can extract rich information about the objects and scenes present, which is complementary to the motion information from clip-level features. The MCN model, for example, uses VGG.

# Reference [19]: Fast R-CNN (Girshick, 2015)

## Key Concepts and Methodology

This paper proposes Faster R-CNN, an object detection framework designed to identify and localize objects within an image. Its key innovation is the Region Proposal Network (RPN), which is a fully convolutional network that shares features with the main detection network. The RPN efficiently predicts a set of object proposals (bounding boxes with an 'objectness' score), which are then classified and refined by the rest of the network.

## Main Contributions and Findings

By integrating region proposal into the main detection network, Faster R-CNN created a unified, end-to-end framework for object detection that was significantly faster and more accurate than its predecessors, enabling near real-time performance.

## Relevance to Primary Survey

Cited as a source for frame-level, object-centric features. NLVL models can use Faster R-CNN to extract features not just for the whole frame, but for specific objects within it. This allows for a more fine-grained understanding of the video content, which is crucial for localizing queries that mention specific objects (e.g., 'the person throwing the red ball'). The SLTA and DORi models are noted to use these features.

# Reference [20]: Skip-Thought Vectors (Kiros et al., 2015)

## Key Concepts and Methodology

This paper introduced Skip-Thought, an unsupervised model for learning generic, distributed sentence representations. The model is trained on a large corpus of text to encode a given sentence such that it can predict the sentences immediately preceding and following it. The resulting sentence encoder can be used as a general-purpose feature extractor for various downstream NLP tasks.

## Main Contributions and Findings

The model provides a powerful way to obtain a single vector representation for an entire sentence that captures its semantic meaning, analogous to how Word2Vec captures word meaning. This allows for semantic similarity comparisons between sentences.

## Relevance to Primary Survey

Cited as a method for extracting a holistic feature vector for the entire language query. While many NLVL models focus on word-level features, the survey notes that some models, like CTRL, use skip-thought vectors to get a global sentence-level representation.

# Reference [21]: Natural Language Processing with Python (Bird et al., 2009)

## Key Concepts and Methodology

This book introduces the Natural Language Toolkit (NLTK), which is a comprehensive suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the English language. It provides easy-to-use interfaces to over 50 corpora and lexical resources, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

## Relevance to Primary Survey

Cited as a foundational tool used in the text feature extraction pipeline of many NLVL models. NLTK is often used for pre-processing steps like tokenizing a sentence into words before they are converted into numerical vectors using embedding models.

# Reference [22]: GloVe: Global Vectors for Word Representation (Pennington et al., 2014)

## Key Concepts and Methodology

This paper introduced GloVe (Global Vectors), an unsupervised learning algorithm for obtaining vector representations (embeddings) for words. GloVe's training is based on aggregated global word-word co-occurrence statistics from a corpus. The model learns word vectors such that their dot product equals the logarithm of their co-occurrence probability, effectively capturing semantic relationships like analogies (e.g., king - man + woman = queen).

## Main Contributions and Findings

GloVe provided an alternative to Word2Vec that is based on matrix factorization of global co-occurrence statistics. Its resulting word embeddings became a standard in the NLP community.

## Relevance to Primary Survey

Cited as a standard, and perhaps the most common, component for text feature extraction in NLVL. The survey highlights that a vast majority of models, including MCN, CTRL, and ACL, use pre-trained GloVe embeddings to convert words in the query sentence into numerical vectors, which is the first step in the language encoding pipeline.

# Reference [23]: Long Short-Term Memory (Hochreiter & Schmidhuber, 1997)

## Key Concepts and Methodology

This seminal paper introduced the Long Short-Term Memory (LSTM) architecture, a type of recurrent neural network (RNN). LSTMs are explicitly designed to avoid the long-term dependency problem (i.e., vanishing and exploding gradients) that affects simple RNNs. They achieve this through a sophisticated cell structure containing input, output, and forget gates that regulate the flow of information, allowing the network to remember information for long periods.

## Relevance to Primary Survey

Cited as a standard model for processing sequential data like text in NLVL. After converting words to embeddings, an LSTM is often used to process the sequence of word vectors to capture contextual and sequential information in the query sentence. Models like CTRL use LSTMs for sentence encoding.

# Reference [24]: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (Chung et al., 2014)

## Key Concepts and Methodology

This paper provides an empirical evaluation of the Gated Recurrent Unit (GRU), a gating mechanism in recurrent neural networks that is similar to an LSTM. A GRU combines the forget and input gates of an LSTM into a single 'update gate' and merges the cell state and hidden state. This results in a simpler architecture with fewer parameters than a standard LSTM.

## Relevance to Primary Survey

Cited alongside LSTM as another common model used in the text feature extraction pipeline of NLVL models. Due to its simpler architecture and comparable performance on many tasks, the GRU is often used as an alternative to the LSTM for processing sequences of word features.

# Reference [25]: Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos (Zhang et al., 2019)

## Key Concepts and Methodology

The CMIN model is a dense locator that enhances the 'TGN-scheme' for NLVL. It introduces a multi-stage cross-modal interaction module to progressively refine the alignment between video and text features. The model first uses a syntactic Graph Convolutional Network (GCN) to capture dependencies in the text query. This is followed by multi-head self-attention and a series of interaction blocks to handle complex sentences and long-range temporal dependencies in the video.

## Relevance to Primary Survey

Presented as a significant and advanced model within the 'dense locators' category. It exemplifies how the influential TGN approach can be extended with more sophisticated interaction mechanisms (like GCNs and multi-stage fusion) to improve performance on complex queries.

# Reference [26]: Localizing Moments in Video with Natural Language (Hendricks et al., 2017)

## Key Concepts and Methodology

This paper introduced the Moment Context Network (MCN), a proposal-based supervised method. The model first generates a pool of candidate video segments (moments) of different temporal scales. For each candidate, it computes the semantic similarity between its visual features and the language query's features. A key aspect is the use of 'moment context' by comparing a candidate with its surrounding video segments to improve localization.

## Main Contributions and Findings

This work is credited with pioneering the systematic study of NLVL. It also introduced the DiDeMo (Distinct Descriptions of Moments in Videos) dataset, which became one of the standard benchmarks in the field.

## Relevance to Primary Survey

Cited as one of the earliest and most foundational studies in the NLVL field. It established the proposal-based paradigm, where the task is framed as ranking a set of pre-defined candidate segments, a strategy adopted by many subsequent papers.

# Reference [27]: Localizing Moments in Video with Temporal Language (Hendricks et al., 2018)

## Key Concepts and Methodology

This paper proposed the Moment Localization with Latent Context (MLLC) model. A key focus of this work is to better model the temporal aspects of the language query itself. It introduces a dataset specifically designed to contain queries with explicit temporal language (e.g., 'the person opens the door \*before\* sitting down'). The MLLC model then learns to ground these temporal phrases in the video.

## Main Contributions and Findings

The paper introduced the TEMPO dataset, which was specifically created to push research towards understanding temporal relationships expressed in language (e.g., 'before', 'after', 'while').

## Relevance to Primary Survey

Presented as a direct follow-up to the authors' previous work (MCN). It highlights a shift in focus from just matching content to explicitly modeling the temporal structure and relationships described in the query sentence.

# Reference [28]: TALL: Temporal Activity Localization via Language Query (Gao et al., 2017)

## Key Concepts and Methodology

This paper introduced the Cross-modal Temporal Regression Localizer (CTRL), also referred to as TALL. It is a pioneering proposal-based supervised model that uses a sliding window approach at multiple scales to generate candidate temporal regions. Its primary innovation is a multi-task loss function that combines an alignment loss (to ensure semantic similarity) and a regression loss (to fine-tune the start and end boundaries of the best-matching window).

## Main Contributions and Findings

This work was one of the first to incorporate a regression objective to explicitly refine window boundaries, which significantly improves localization precision over methods that only rely on ranking. It also introduced the Charades-STA dataset, another standard benchmark for NLVL.

## Relevance to Primary Survey

Highlighted as one of the pioneering studies in NLVL. Its hybrid loss function design, which considers both semantic alignment and boundary regression, proved to be highly effective and was adopted in many subsequent proposal-based models.

# Reference [29]: Attentive Moment Retrieval in Videos (Liu et al., 2018)

## Key Concepts and Methodology

This paper proposed a proposal-based method, ACRN, that leverages an attention mechanism to improve cross-modal alignment. Specifically, the sentence feature is used as a query to 'attend' to the most relevant parts of the video segment's features. This allows the model to dynamically weigh different parts of the video feature representation, focusing on the elements that are most relevant to the language query. The model also explores feature interaction via outer products.

## Relevance to Primary Survey

Cited as an early and influential work that successfully demonstrated the effectiveness of applying attention mechanisms for NLVL. It showed that attention could be used to mine finer-grained matching relationships between the video and text modalities within the proposal-based framework.

# Reference [30]: Cross-Modal Moment Localization in Videos (Liu et al., 2018)

## Key Concepts and Methodology

This paper, a companion to the ACRN work, also proposes a proposal-based method that uses an attention mechanism. However, in this formulation, the attention is reversed: the video segment's features are used to guide the sentence features. This means the model learns to focus on the most relevant words or phrases in the query that correspond to the visual content of the candidate moment.

## Relevance to Primary Survey

Paired with its companion paper [29], this work is highlighted as another successful and early application of attention for cross-modal interaction in NLVL. The two papers together explore the duality of attention, showing its effectiveness in both directions (text-to-video and video-to-text).

# Reference [31]: Multi-Modal Circulant Fusion for Video-to-Language and Backward (Wu & Han, 2018)

## Key Concepts and Methodology

This paper proposed Multi-modal Circulant Fusion (MCF), a sophisticated method for cross-modal interaction. It constructs a circulant matrix from one modality's feature vector and uses it to perform a convolution-like operation with the other modality's vector. This allows for rich, element-level interactions between all parts of the feature vectors.

## Limitations and Broader Context

The primary drawback of this approach is its high computational cost. The circulant matrix operation significantly increases the number of computations required, which can make the model slow to train and deploy.

## Relevance to Primary Survey

Cited as an example of a sophisticated and powerful design for cross-modal interaction within the proposal-based framework. It represents an exploration into more expressive, albeit computationally expensive, fusion techniques.

# Reference [32]: Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos (Liu et al., 2018)

## Key Concepts and Methodology

This paper introduced Temporal Modular Networks (TMN), which treat the language query not as a flat sequence of words but as a structured sentence. It parses the query into a syntactic tree (e.g., subject-verb-object) and composes the final sentence representation from its grammatical components. This allows the model to better understand complex queries describing compositional activities.

## Relevance to Primary Survey

Cited as a method that moves beyond simple bag-of-words or sequential sentence representations. By explicitly modeling the grammatical structure of the query, it aims to achieve better localization of complex, multi-part activities.

# Reference [33]: Exploiting Temporal Relationships in Video Moment Localization with Natural Language (Zhang et al., 2019)

## Key Concepts and Methodology

This work proposed the Temporal Compositional Modular Network (TMCN). Similar to TMN, it leverages the grammatical structure of the sentence by parsing it into a tree. The model then uses this structure to guide the composition of video features, aiming to match the compositional nature of the language query with the visual evidence.

## Relevance to Primary Survey

Grouped with TMN as another example of a proposal-based method that focuses on exploiting the internal linguistic structure of the language query to improve localization accuracy for complex events.

# Reference [34]: MAC: Mining Activity Concepts for Language-Based Temporal Localization (Ge et al., 2019)

## Key Concepts and Methodology

This paper proposed the Activity Concepts based Localizer (ACL). Instead of matching raw features, ACL first mines explicit 'activity concepts' (such as verb-object pairs) from both the sentence and the video. For instance, it might identify 'throwing a ball' as a key concept. The final localization is then performed by aligning these high-level semantic concepts.

## Main Contributions and Findings

The model incorporates classification layer features from C3D (as visual concepts) and verb-object pair features from GloVe embeddings (as textual concepts) to create a more interpretable and robust matching process.

## Relevance to Primary Survey

Presented as a method that goes beyond general, holistic feature matching. By focusing on specific, mined 'activity concepts,' it guides the localization process with more explicit semantic cues.

# Reference [35]: Semantic Proposal for Activity Localization in Videos via Sentence Query (Chen & Jiang, 2019)

## Key Concepts and Methodology

This paper proposes a unique method where the language query is first transformed into a 'visual semantic' representation. It uses a pre-defined lexicon of common words and learns to map the sentence to a combination of these visual concepts. These concepts are then used to generate and score candidate video segments, effectively bridging the modality gap by translating language into a visual-like space.

## Relevance to Primary Survey

Cited as an example of a proposal-based method that employs a novel strategy for cross-modal matching. The idea of transforming language into a 'visual semantic concept' representation offers an interesting alternative to direct feature fusion.

# Reference [36]: Cross-Modal Video Moment Retrieval with Spatial and Language-Temporal Attention (Jiang et al., 2019)

## Key Concepts and Methodology

This paper proposed the Spatial and Language-Temporal Attention (SLTA) model. A key innovation is the incorporation of fine-grained, object-level visual features extracted using a Faster R-CNN detector. The model then uses attention mechanisms to interact the language query with both the motion features of the video and the object features within the frames.

## Main Contributions and Findings

By integrating object-level information, the model can ground parts of the query (e.g., 'the man in the blue shirt') to specific objects, leading to more precise and fine-grained matching.

## Relevance to Primary Survey

Cited as a key example of leveraging detailed, object-level information from video frames to enhance localization accuracy. This moves beyond coarse, whole-frame features and allows for a deeper understanding of the scene.

# Reference [37]: Find and Focus: Retrieve and Localize Video Events with Natural Language Queries (Shao et al., 2018)

## Key Concepts and Methodology

This paper introduces a 'correlation-based proposal' method. Instead of using naive, fixed-size sliding windows, it first computes a correlation score between each video clip and the entire sentence query. The regions with high correlation are then used to generate a much smaller and more relevant set of candidate windows for final ranking.

## Limitations and Broader Context

This work was motivated by the inefficiency and inaccuracy of naive proposal methods that use fixed window sizes, which often fail to align well with the true event boundaries.

## Relevance to Primary Survey

Cited as an important work that improved the efficiency and effectiveness of the proposal generation step. By using query-video correlation to create proposals, it generates more meaningful candidate segments from the outset.

# Reference [38]: Multilevel Language and Vision Integration for Text-to-Clip Retrieval (Xu et al., 2019)

## Key Concepts and Methodology

This paper proposed the Query-guided Segment Proposal Network (QSPN), which also falls under the 'correlation-based proposal' category. Additionally, it explored using video captioning as an auxiliary task during training, forcing the model to generate a description for a video segment, which helps regularize the learned features. It also systematically studied the effect of feature fusion at different stages.

## Main Contributions and Findings

The paper provided experimental evidence that earlier interaction between the features of the two modalities generally leads to better performance than late fusion.

## Relevance to Primary Survey

Cited for its dual contributions: advancing correlation-based proposal generation and for its empirical finding on the importance of early-stage feature interaction, which influenced subsequent architectural designs.

# Reference [39]: Interaction-Integrated Network for Natural Language Moment Localization (Ning et al., 2021)

## Key Concepts and Methodology

This paper proposed an Interaction-Integrated Network (I2N) designed to capture long-range contextual information in the video. It achieves this by stacking 'Interaction-Integrated Cells,' which are complex modules containing multiple types of feature interactions like dot product, summation, and attention, allowing for a comprehensive fusion of video and text features.

## Limitations and Broader Context

This work was motivated by the information loss that can occur in traditional windowing approaches, especially when the proposal windows do not perfectly align with the target moment.

## Relevance to Primary Survey

Cited as an advanced proposal-based method that uses a complex, stacked interaction module to better model the video's structure and the query's relationship to it.

# Reference [40]: Temporally Grounding Natural Sentence in Video (Chen et al., 2018)

## Key Concepts and Methodology

This paper designed the Temporal Ground Net (TGN), a 'dense locator' method. Unlike proposal-based methods, TGN avoids generating candidate windows beforehand. Instead, at each time step in the video, it predicts a set of candidate moments of different, pre-defined durations that all \*end\* at that time step. This is achieved by using an LSTM over the video features, where the output at each step is expected to contain multi-scale temporal information.

## Main Contributions and Findings

It introduced the highly influential 'TGN-scheme,' which posits that video features at each timestep after being processed by an RNN can implicitly contain temporal information at multiple scales. This counter-intuitive but effective scheme was adopted by many subsequent dense locator models.

## Relevance to Primary Survey

Cited as the pioneering work for the 'Dense Locators' category. Its novel scheme for predicting multiple windows at each timestep provided a new and efficient paradigm for NLVL, moving away from the two-stage proposal-and-rank approach.

# Reference [41]: Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction (Lin et al., 2020)

## Key Concepts and Methodology

This paper is a direct extension of the CMIN model [25]. In addition to the multi-stage cross-modal interaction, it introduces a query reconstruction loss. After the model has fused the video and text features to make a prediction, a decoder attempts to reconstruct the original text query from the fused representation. This acts as a regularization technique.

## Relevance to Primary Survey

Cited as a follow-up to the CMIN paper. By adding a query reconstruction task, the model is forced to preserve more of the semantic information from the language query in its internal representations, which further improved performance.

# Reference [42]: DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization (Lu et al., 2019)

## Key Concepts and Methodology

This paper follows the TGN-scheme and frames the problem as a 'bottom-up' approach, contrasting with 'top-down' proposal-based methods. It uses QANet as its backbone network and employs three separate prediction heads trained with three distinct losses: a classification loss for moment presence, a regression loss for boundary refinement, and a scoring loss for ranking.

## Main Contributions and Findings

The model is presented as a more efficient bottom-up alternative to proposal-based methods, as it avoids the costly generation and processing of thousands of candidate windows.

## Relevance to Primary Survey

Cited as another influential 'Dense Locator' model that successfully adopted and built upon the TGN-scheme, demonstrating its effectiveness and flexibility.

# Reference [43]: QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension (Yu et al., 2018)

## Key Concepts and Methodology

QANet is a high-performance Question Answering model. Its architecture is notable for its exclusive use of convolution and self-attention, completely eschewing recurrent networks like LSTMs. It uses local convolutional layers to model local interactions and global self-attention layers to model long-range dependencies, resulting in a much faster and more parallelizable architecture.

## Relevance to Primary Survey

Cited as the backbone network architecture used by the NLVL models DEBUG and VSLNet. This highlights a key trend in the field: successfully adapting state-of-the-art architectures from other NLP domains, like question answering, for the NLVL task.

# Reference [44]: Rethinking the Bottom-Up Framework for Query-Based Video Localization (Chen et al., 2020)

## Key Concepts and Methodology

This paper proposed Graph FPN with Dense Predictions (GDP), which improves upon the DEBUG model. It introduces two key enhancements: a graph convolution module to better model relationships between video clips and a Feature Pyramid Network (FPN) to handle features at multiple temporal scales more effectively.

## Relevance to Primary Survey

Cited as an incremental but important improvement over the DEBUG model. It showcases the benefits of incorporating more sophisticated architectural components, like graph networks and feature pyramids, into the dense locator framework.

# Reference [45]: Dense Regression Network for Video Grounding (Zeng et al., 2020)

## Key Concepts and Methodology

This paper proposed the Dense Regression Network (DRN), which adopts the dense prediction head idea from other models. Its key feature is a dedicated IoU (Intersection over Union) regression head. During training, this head is explicitly trained to predict the IoU between a predicted moment and the ground truth, forcing the model to directly optimize for localization quality.

## Relevance to Primary Survey

Cited as an example of a 'Dense Locator' model that specifically focuses on improving localization quality. By adding a module that is explicitly aware of the final evaluation metric (IoU), it encourages the network to produce more accurate boundaries.

# Reference [46]: Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization (Liu et al., 2020)

## Key Concepts and Methodology

This paper proposed CSMGAN, which uses a TGN-like scheme but with a sophisticated graph-based interaction module. It constructs two types of graphs: a Cross-Modal relation Graph (CMG) to model interactions between video and text features, and a Self-Modal relation Graph (SMG) to model intra-modal relationships (e.g., context within the video). It also processes text at multiple granularities (word, phrase, sentence).

## Main Contributions and Findings

The model uses multiple layers of graph attention to achieve deep and comprehensive feature interaction both across and within the two modalities.

## Relevance to Primary Survey

Cited as a highly sophisticated 'Dense Locator' that leverages the power of graph neural networks to achieve state-of-the-art performance by modeling a rich web of inter- and intra-modal relationships.

# Reference [47]: Fine-Grained Iterative Attention Network for Temporal Language Localization in Videos (Qu et al., 2020)

## Key Concepts and Methodology

This paper proposed the Fine-grained Iterative Attention Network (FIAN). Unlike TGN-scheme models that predict windows ending at each time step, FIAN uses a more intuitive approach of applying true sliding windows of different scales over the fused video-text features. It also uses an 'information gate' that allows the modalities to iteratively guide each other's representations.

## Main Contributions and Findings

The authors argue that this approach is more interpretable than the TGN-scheme and it achieved strong experimental results, demonstrating a viable alternative to the dominant dense locator design.

## Relevance to Primary Survey

Cited as a 'Dense Locator' model that uses a more intuitive and interpretable multi-scale localization strategy. It shows that there are alternative ways to design dense locators beyond the influential TGN-scheme.

# Reference [48]: Attention on Attention for Image Captioning (Huang et al., 2019)

## Key Concepts and Methodology

This paper introduced an 'Attention on Attention' module for image captioning. It includes an 'information gate' that uses the output of the attention module itself to decide how much of the visual context should be passed on to the language generation model. This helps in filtering out irrelevant information and improving caption quality.

## Relevance to Primary Survey

Cited as the source of the 'information gate' mechanism that was adapted for use in the FIAN model [47]. This is a clear example of the cross-pollination of architectural ideas from the image captioning domain to the NLVL domain.

# Reference [49]: MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment (Zhang et al., 2019)

## Key Concepts and Methodology

This paper proposed MAN, a 'Dense Locator' model that uses a multi-scale locator strategy. Its novelty lies in two components: it uses dynamic convolution (where the filter weights are generated based on the input) for modal fusion, and an Iterative Graph Adjustment Network to refine the relationships between candidate moments and the query.

## Relevance to Primary Survey

Cited as another example of a 'Dense Locator' model that uses a multi-scale localization strategy. It also successfully experimented with using features from the TAN model instead of VGG, highlighting the impact of the underlying feature extractor on final performance.

# Reference [50]: Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos (Yuan et al., 2019)

## Key Concepts and Methodology

This paper proposed a semantic conditioned dynamic modulation (SCDM) mechanism. The core idea is that the semantics of the language query should dynamically modulate the process of temporal convolution over the video features. In essence, the sentence 'conditions' how the video is perceived by the model. It also employs a multi-scale locator strategy.

## Relevance to Primary Survey

Cited as another example of a 'Dense Locator' with a multi-scale strategy, similar to MAN and FIAN. Its unique contribution is the idea of dynamically altering the video processing pipeline based on the language query.

# Reference [51]: TAN: Temporal Aggregation Network for Dense Multi-Label Action Recognition (Dai et al., 2019)

## Key Concepts and Methodology

This paper introduces TAN, a network designed for efficient temporal aggregation in action recognition. It proposes a lightweight temporal aggregation module that can be flexibly inserted into existing 2D CNNs to enable temporal modeling without the high computational cost of 3D convolutions.

## Relevance to Primary Survey

Cited as the source for the video features used in the MAN model [49]. The survey notes that MAN's performance improved when using TAN features compared to VGG features, which underscores the critical importance of using powerful and task-relevant feature extractors.

# Reference [52]: Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language (Zhang et al., 2020)

## Key Concepts and Methodology

This work proposed a unique 'Dense Locator' approach by constructing a two-dimensional temporal feature map. In this 2D map, the coordinate (i, j) represents the video segment spanning from clip i to clip j. The model then computes a similarity score between this entire 2D map and the query features, effectively turning the problem into a 2D image-like matching task.

## Limitations and Broader Context

This 2D map representation is highly computationally and memory intensive, as its size is quadratic with respect to the video length. The authors had to employ tricks like sampling to make the approach feasible.

## Relevance to Primary Survey

Cited as a novel 'Dense Locator' approach that conceptualizes the problem in a completely different way (a 2D temporal map vs. a 1D sequence). It demonstrates the creativity in the field but also highlights the computational trade-offs involved.

# Reference [53]: Dual Path Interaction Network for Video Moment Localization (Wang et al., 2020)

## Key Concepts and Methodology

This paper proposed the Dual Path Interaction Network (DPIN), which processes information at both the frame-level (individual frames) and the candidate-level (video segments) simultaneously. It uses a 'Semantically Conditioned Interaction' module to allow information to flow and transform between these two parallel paths.

## Main Contributions and Findings

The core motivation is that these two levels of feature representation are complementary. Frame-level features offer fine-grained detail, while candidate-level features provide temporal context. Modeling their interaction explicitly leads to better localization.

## Relevance to Primary Survey

Cited as an innovative 'Dense Locator' model that explicitly models the interaction between different levels of feature representation (frames vs. segments) to gain a more holistic understanding of the video.

# Reference [54]: To Find Where You Talk: Temporal Sentence Localization in Video with Attention Based Location Regression (Yuan et al., 2019)

## Key Concepts and Methodology

This paper constructed the Attention Based Location Regression (ABLR) model, a 'Single-Shot' method. In this approach, after computing cross-modal attention weights over the video, the model does not use the attended features for prediction. Instead, it directly uses the attention weights themselves as features, feeding them into a simple MLP to directly regress the start and end boundaries of the moment.

## Main Contributions and Findings

The experimental results showed that, for this architecture, predicting directly from the attention distribution was more effective than predicting from the feature vectors produced by that attention.

## Relevance to Primary Survey

Cited as a key example of a 'Single-Shot' localization model. This category of models aims to directly predict the final boundaries in one go, without intermediate proposal or dense prediction steps.

# Reference [55]: Local-Global Video-Text Interactions for Temporal Grounding (Mun et al., 2020)

## Key Concepts and Methodology

This high-performing single-shot model achieves its results by effectively combining several established and powerful mechanisms. It uses position embeddings to encode temporal location, the Hadamard product for efficient feature fusion, and non-local modules to capture long-range dependencies across the entire video. It also introduces a distinct query attention loss to focus on the most important words in the sentence.

## Relevance to Primary Survey

Cited as a strong 'Single-Shot' model that demonstrates how thoughtfully combining multiple well-established techniques can lead to state-of-the-art performance. It serves as an example of successful engineering in the field.

# Reference [56]: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2018)

## Key Concepts and Methodology

BERT is a landmark language representation model based on the Transformer architecture. It is pre-trained on a massive amount of unlabeled text using two novel tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). This allows it to learn deep bidirectional representations of text, meaning it understands the context of a word from both left and right.

## Relevance to Primary Survey

Cited as the source of the 'position embedding' concept, which was adapted for use in the LGI model [55]. This shows the profound influence of large language models and the Transformer architecture on the design of NLVL models.

# Reference [57]: Hadamard Product for Low-Rank Bilinear Pooling (Kim et al., 2016)

## Key Concepts and Methodology

This paper explores the use of the Hadamard product (element-wise multiplication) as an efficient technique for feature fusion. It is presented as a computationally cheaper alternative to other methods like outer product for achieving bilinear pooling, which can capture complex second-order interactions between features.

## Relevance to Primary Survey

Cited as the source of the 'Hadamard product' fusion technique used in the LGI model [55]. It highlights how NLVL models incorporate specific mathematical operations and techniques from the broader machine learning literature.

# Reference [58]: Non-Local Neural Networks (Wang et al., 2018)

## Key Concepts and Methodology

This paper introduced non-local modules, which are designed to capture long-range dependencies. A non-local operation computes the response at a position as a weighted sum of the features at all positions in the input. This is conceptually similar to self-attention and allows the network to model dependencies between distant points in space or time directly.

## Relevance to Primary Survey

Cited as the source of the 'non-local modules' used in the LGI model [55] to capture a broader temporal context beyond what an RNN or local convolution could see.

# Reference [59]: Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos (Chen et al., 2020)

## Key Concepts and Methodology

This paper proposed the Channel-Gated Modality Interaction model. A key aspect of this single-shot model is its explicit use of multiple video modalities, such as visual appearance, motion, and audio. It uses a gating mechanism to control the flow of information and perform pair-wise interaction across all these modalities to find the best alignment with the text query.

## Relevance to Primary Survey

Cited as a 'Single-Shot' model that explicitly explores using multi-modal video information. Most NLVL models rely solely on visual features, but this work shows the potential benefit of incorporating other data streams like motion and audio.

# Reference [60]: Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language (Chen & Jiang, 2020)

## Key Concepts and Methodology

This paper proposed the Hierarchical Visual-Textual Graph (HVTG). It argues that past work neglected the importance of objects. The model constructs a hierarchical graph to perform visual-textual interaction at two levels: the object level and the frame (channel) level. It also introduces a feature normalization loss to help the model predict more reliable relevance scores.

## Relevance to Primary Survey

Cited as a 'Single-Shot' model that strongly emphasizes the role of objects in localization. It uses a hierarchical graph structure to explicitly model object-word and frame-word interactions, combined with a novel loss function.

# Reference [61]: Ring Loss: Convex Feature Normalization for Face Recognition (Zheng et al., 2018)

## Key Concepts and Methodology

This paper introduces Ring Loss, a feature normalization technique designed for face recognition. It encourages the learned deep features to have a fixed magnitude (i.e., lie on a hypersphere) while still being discriminative. This is achieved by adding a loss term that penalizes the deviation of feature norms from a learned target value.

## Relevance to Primary Survey

Cited as the source of the 'feature normalization loss' used in the HVTG model [60]. This is another example of NLVL research borrowing specific loss function designs from other computer vision domains like face recognition.

# Reference [62]: Attention-Aware Deep Reinforcement Learning for Video Face Recognition (Rao et al., 2017)

## Key Concepts and Methodology

This paper applies deep reinforcement learning (DRL) to the task of video face recognition. An RL agent learns a policy to selectively process frames or regions within a video to efficiently identify a person, focusing its computational efforts on the most informative parts of the video.

## Relevance to Primary Survey

Cited as an example of the successful application of DRL in the broader field of video understanding. This success serves as a motivation and justification for applying similar RL-based approaches to the NLVL task.

# Reference [63]: Reinforcement Cutting-Agent Learning for Video Object Segmentation (Han et al., 2018)

## Key Concepts and Methodology

This work uses reinforcement learning for the task of video object segmentation. An agent learns a policy to 'cut out' an object from a video sequence, making sequential decisions to refine the segmentation mask over time. This interactive, decision-making framework is well-suited for the task.

## Relevance to Primary Survey

Cited as part of the body of work demonstrating that reinforcement learning has developed rapidly and proven effective for various video understanding tasks, making it a viable and promising approach for NLVL.

# Reference [64]: Deep Reinforcement Learning for General Video Game AI (Torrado et al., 2018)

## Key Concepts and Methodology

This paper provides a survey and exploration of using deep reinforcement learning to create general artificial intelligence agents that can learn to play a wide variety of video games. This domain requires agents to make sequential decisions in complex, dynamic environments to achieve a goal.

## Relevance to Primary Survey

Cited to show the breadth and power of reinforcement learning applications. If RL can handle complex, interactive environments like video games, it is plausible that it can also handle the sequential decision-making process required for NLVL.

# Reference [65]: Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos (He et al., 2019)

## Key Concepts and Methodology

This paper proposed the RWM model, which frames NLVL as a reinforcement learning problem. The RL agent starts with an initial candidate window in the video. At each step, it observes the current state (the window's features and the query) and chooses an action from a set like {move left, move right, shrink, expand, stop}. It receives a reward based on how much its action improved the window's IoU with the ground truth.

## Main Contributions and Findings

This intuitive 'move and resize' strategy is noted to be highly interpretable and achieves good performance by actively searching for the correct moment.

## Relevance to Primary Survey

Cited as a key example of the Reinforcement Learning approach to NLVL. It directly models the localization task as a sequential decision-making process.

# Reference [66]: Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model (Wang et al., 2019)

## Key Concepts and Methodology

This paper proposed the SM-RL model. It also uses reinforcement learning, but with a different state and action design. The agent selectively observes a sequence of video fragments to gather information, and based on this sequence, it predicts the final start and end boundaries. Both this model and RWM use an RNN to maintain a memory of the agent's state at each timestep.

## Relevance to Primary Survey

Cited as another primary example of the Reinforcement Learning paradigm for NLVL. It contrasts with the RWM model in its observation strategy, showcasing that different RL formulations are possible for the same core problem.

# Reference [67]: Tree-Structured Policy Based Progressive Reinforcement Learning for Temporally Language Grounding in Video (Wu et al., 2020)

## Key Concepts and Methodology

This paper proposed the Tree-Structured Policy based Progressive Reinforcement Learning (TSP-PRL) model. It enhances the RL approach with a hierarchical policy. The agent first chooses a coarse action type (e.g., 'shift boundary') from a root-level policy, and then chooses a fine-grained action magnitude (e.g., 'shift by 2 frames') from a corresponding leaf-level policy. This allows for more precise, fine-tuned adjustments.

## Relevance to Primary Survey

Cited as an advanced Reinforcement Learning model that uses a hierarchical policy structure. This enables more nuanced and controlled adjustments to the predicted boundaries compared to a flat action space.

# Reference [68]: Tripping Through Time: Efficient Localization of Activities in Videos (Hahn et al., 2020)

## Key Concepts and Methodology

This RL-based paper, called TripNet, introduces a gated attention mechanism to align the text and video modalities. The RL agent learns to navigate through the video ('trip through time'), and the gated attention helps it focus on the most relevant parts of the query and video at each step, making the localization process more efficient.

## Main Contributions and Findings

The model is noted for achieving faster localization speeds than many traditional, non-RL approaches.

## Relevance to Primary Survey

Cited as a Reinforcement Learning model that successfully incorporates gated attention for more efficient and effective localization.

# Reference [69]: Gated-Attention Readers for Text Comprehension (Dhingra et al., 2016)

## Key Concepts and Methodology

This paper introduces the concept of gated attention for text comprehension tasks. In this mechanism, the attention weights are modulated by a gating function that is conditioned on the query. This allows the model to learn to control the flow of information from the context to the query representation.

## Relevance to Primary Survey

Cited as the source of the 'gated attention' mechanism that was adapted for use in the TripNet model [68], showing how specific attention variants are borrowed from the NLP literature.

# Reference [70]: Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization (Cao et al., 2020)

## Key Concepts and Methodology

This paper proposed an adversarial learning paradigm that combines reinforcement learning with moment ranking. It trains a generator (the RL agent) to propose good candidate moments, and a discriminator to distinguish these proposed moments from ground-truth moments. The generator is rewarded for fooling the discriminator, creating an adversarial game that pushes the RL agent to find better moments.

## Main Contributions and Findings

This approach helps the model learn the subtle differences between relevant and irrelevant moments within the same video, improving its discriminative ability.

## Relevance to Primary Survey

Cited as a Reinforcement Learning model that uniquely incorporates adversarial learning (GANs) to enhance its training signal and improve performance.

# Reference [71]: STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization (Cao et al., 2020)

## Key Concepts and Methodology

The STRONG model extends the RL paradigm for NLVL into the spatial dimension. The RL agent learns a policy not only to select the correct temporal segment but also to select the most important spatial regions (bounding boxes) within the frames of that segment. It can adjust its regions of interest in both space and time.

## Main Contributions and Findings

By learning to focus on relevant spatial areas, the model can filter out irrelevant background distractions, leading to more robust localization, especially for queries that mention specific objects.

## Relevance to Primary Survey

Cited as a Reinforcement Learning model that expands the problem from purely temporal localization to spatio-temporal localization, making it a more comprehensive approach.

# Reference [72]: Localizing Natural Language in Videos (Chen et al., 2019)

## Key Concepts and Methodology

This paper introduced a 'boundary-aware' approach, which reframes the NLVL problem. Instead of predicting a segment (start and end), it predicts the probability of each frame being a start boundary and the probability of it being an end boundary. It designed a 'self interactor' module with extensive attention operations to capture the rich contextual information needed to make these boundary decisions.

## Main Contributions and Findings

This method transforms the localization problem from finding a segment to finding two points, which completely avoids the need for generating candidate windows or proposals.

## Relevance to Primary Survey

Cited as one of the pioneering works in the 'Boundary Aware' category. It established a new and efficient localization strategy that directly targets the start and end points of the moment.

# Reference [73]: ExCL: Extractive Clip Localization Using Natural Language Descriptions (Ghosh et al., 2019)

## Key Concepts and Methodology

This paper also used a 'boundary-aware' approach, similar to [72]. It frames the task as 'extractive clip localization' and explores the structure of three simple, independent localizers for the start, middle, and end of the event. This simplified and modular design helps to avoid a large number of computations.

## Relevance to Primary Survey

Cited alongside [72] as another foundational paper for the 'Boundary Aware' strategy. It highlights that this approach can be implemented with a focus on computational efficiency.

# Reference [74]: Temporally Grounding Language Queries in Videos by Contextual Boundary-Aware Prediction (Wang et al., 2020)

## Key Concepts and Methodology

This paper, following the boundary-aware paradigm, designed a specific contextual integration module. This module is dedicated to capturing rich contextual information from both before and after a specific time step to help the model accurately predict if that time step is a moment boundary. It explicitly focuses on modeling the context needed for boundary decisions.

## Relevance to Primary Survey

Cited as a 'Boundary Aware' model that makes a targeted contribution by designing a specialized module to better capture context, which is the most critical element for accurate boundary prediction.

# Reference [75]: Proposal-Free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention (Rodriguez-Opazo et al., 2020)

## Key Concepts and Methodology

This work (PFGA) uses a boundary-aware approach with a simple and elegant structure. It employs a guided attention filter to help the model focus on relevant features and uses a Kullback-Leibler (KL) divergence loss to match the predicted boundary probability distributions with a target distribution. This combination achieves good performance without complex architectural components.

## Relevance to Primary Survey

Cited as a 'Boundary Aware' model that demonstrates strong results can be achieved with a relatively simple and efficient architecture, provided the right loss function and attention mechanisms are used.

# Reference [76]: Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models (Hershey & Olsen, 2007)

## Key Concepts and Methodology

This paper is a theoretical work that discusses various methods for approximating the Kullback-Leibler (KL) divergence. The KL divergence is a measure of how one probability distribution differs from a second, reference probability distribution. It is a core concept in information theory and statistics.

## Relevance to Primary Survey

Cited as the source of the 'Kullback-Leibler divergence' loss used in the PFGA model [75]. This shows how NLVL research draws upon fundamental concepts from information theory to design novel loss functions.

# Reference [77]: DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video (Rodriguez-Opazo et al., 2021)

## Key Concepts and Methodology

This is a follow-up work to PFGA [75]. It enhances the boundary-aware approach by designing a spatial-temporal graph algorithm. This graph models relationships between objects over time and helps to transfer information from the language query to the video feature generation process itself. It uses a combination of fine-grained Faster-RCNN object features and I3D motion features.

## Relevance to Primary Survey

Cited as an advanced 'Boundary Aware' model that makes significant strides by incorporating spatial-temporal graphs and explicit object information, leading to state-of-the-art performance.

# Reference [78]: Span-Based Localizing Network for Natural Language Video Localization (Zhang et al., 2020)

## Key Concepts and Methodology

This paper explicitly frames the NLVL task as a question-answering problem, where the 'question' is the language query and the 'answer' is a temporal 'span' in the video. It builds on a QA framework (QANet) and adds a query-guided highlighting (QGH) strategy to help the model focus on the most relevant video segments before predicting the final answer span.

## Relevance to Primary Survey

Cited as a 'Boundary Aware' model that provides a clear conceptual link between NLVL and extractive question answering. This framing allows it to leverage powerful and efficient architectures from the QA domain.

# Reference [79]: Weakly Supervised Dense Event Captioning in Videos (Duan et al., 2018)

## Key Concepts and Methodology

This paper proposed WS-DEC, a weakly-supervised method that tackles both NLVL and dense video captioning simultaneously. It uses an iterative training procedure that leverages the dual nature of these tasks. In one step, it uses estimated locations to train the captioning model, and in the next step, it uses the generated captions to refine the location estimates. This process repeats, allowing the two tasks to mutually improve each other.

## Relevance to Primary Survey

Cited as an early and innovative weakly-supervised method. Its key contribution is the use of a dual-task, iterative training approach to overcome the lack of precise temporal annotations.

# Reference [80]: Iterative Approximation of Fixed Points of Lipschitzian Strictly Pseudocontractive Mappings (Chidume, 1987)

## Key Concepts and Methodology

This is a highly theoretical paper from the field of pure mathematics. It deals with iterative methods for finding fixed points of certain types of mathematical mappings, which is a fundamental concept in functional analysis.

## Relevance to Primary Survey

Cited as the inspiration for the 'iterative training' methodology used in the WS-DEC model [79]. This is a fascinating example of how ideas from pure mathematics can, decades later, inspire algorithms in applied fields like machine learning.

# Reference [81]: Weakly Supervised Video Moment Retrieval from Text Queries (Mithun et al., 2019)

## Key Concepts and Methodology

This paper proposed Text-Guided Attention (TGA), a foundational weakly-supervised method. The approach is straightforward: the video is pre-segmented into a large number of candidate segments at multiple scales. The model then uses Multiple Instance Learning (MIL), where the video is a 'bag' of segments and the sentence is the label. The model is trained to predict a high score for at least one segment in the bag, implicitly identifying the correct moment.

## Relevance to Primary Survey

Cited as a foundational weakly-supervised method that uses a simple and direct strategy of pre-segmenting the video and applying a standard MIL framework to solve the problem.

# Reference [82]: WSLLN: Weakly Supervised Natural Language Localization Networks (Gao et al., 2019)

## Key Concepts and Methodology

This weakly-supervised paper builds directly on the Multiple Instance Learning (MIL) approach. To improve training, it introduces a pseudo-labeling strategy. Within a video 'bag,' it selects the single clip with the highest score as a pseudo-positive example and treats others as negatives. This helps to better distinguish between relevant and irrelevant segments from the same video, reinforcing intra-video differences.

## Relevance to Primary Survey

Cited as a direct improvement over simpler MIL-based approaches like TGA [81]. The introduction of a pseudo-labeling strategy provides a stronger training signal in the weakly-supervised setting.

# Reference [83]: Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos (Wu et al., 2020)

## Key Concepts and Methodology

This paper applies a reinforcement learning (RL) framework to solve the weakly-supervised NLVL task. The RL agent learns a policy to select candidate moments. The key challenge is defining the reward, as there is no ground-truth moment. The reward is instead based on the overall video-sentence similarity score, encouraging the agent to find segments that contribute most to this score.

## Limitations and Broader Context

The motivation is to overcome the drawbacks of naive proposal-based approaches, such as inefficiency and the risk of the model learning from many inaccurately aligned negative examples.

## Relevance to Primary Survey

Cited as the first work to apply reinforcement learning specifically to the \*weakly-supervised\* NLVL setting, demonstrating the flexibility of the RL paradigm.

# Reference [84]: Reinforcement Learning: An Introduction (Sutton & Barto, 2018)

## Key Concepts and Methodology

This is the foundational and most widely cited textbook on reinforcement learning. It comprehensively covers the core concepts of the field, including Markov Decision Processes, dynamic programming, Monte Carlo methods, temporal-difference learning (like Q-learning and SARSA), and policy gradient methods.

## Relevance to Primary Survey

Cited as the source of the fundamental reinforcement learning principles and algorithms that were applied in the weakly-supervised RL model [83].

# Reference [85]: Weakly-Supervised Video Moment Retrieval via Semantic Completion Network (Lin et al., 2020)

## Key Concepts and Methodology

This weakly-supervised paper uses a sentence reconstruction task to guide learning. It first pre-selects a few candidate windows based on similarity. It then uses an Encoder-Decoder structure to try and reconstruct a masked version of the query sentence from the features of the selected window. A successful reconstruction implies a high degree of semantic consistency between the sentence and the video segment.

## Relevance to Primary Survey

Cited as a weakly-supervised method that uses a clever, self-supervised sentence reconstruction task to enforce semantic consistency, providing a training signal without requiring precise temporal annotations.

# Reference [86]: Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos (Zhang et al., 2020)

## Key Concepts and Methodology

This weakly-supervised paper designed a Regularized Two-Branch Proposal Network, building on the 2D-TAN architecture. One branch focuses on identifying the best matching moment within a video (intra-video), while the other branch focuses on contrasting it with moments from other videos (inter-video). It uses a language-aware filter and regularization to enhance the contrast between positive and negative samples.

## Relevance to Primary Survey

Cited as a weakly-supervised method that uses a specialized two-branch network architecture and regularization techniques to explicitly enhance the contrast between positive and negative samples, which is a key challenge in MIL-based approaches.

# Reference [87]: VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval (Ma et al., 2020)

## Key Concepts and Methodology

This paper proposed VLANet to address specific shortcomings in prior weakly-supervised methods, such as the inefficiency of dense candidate windows and the use of coarse, global query representations. It intelligently selects a sparse set of candidate windows and then uses a bidirectional attention mechanism to achieve fine-grained alignment between the query and these candidates.

## Relevance to Primary Survey

Cited as a weakly-supervised method that was specifically designed to solve identified problems in previous works in this category, representing a more targeted approach to improving the state-of-the-art.

# Reference [88]: Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding (Zhang et al., 2020)

## Key Concepts and Methodology

This paper proposed a novel counterfactual contrastive learning framework. To improve the contrast between positive and negative samples, it generates 'counterfactual' negative examples by applying transformations at the feature, interaction, and relation levels. This creates more challenging negative examples for the model to learn from.

## Main Contributions and Findings

It also innovatively uses gradient information (inspired by Grad-CAM) as a factor when selecting positive and negative samples for the contrastive learning process.

## Relevance to Primary Survey

Cited as an advanced weakly-supervised method that introduces a novel counterfactual learning paradigm to significantly improve the training process in the absence of strong supervision.

# Reference [89]: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization (Selvaraju et al., 2017)

## Key Concepts and Methodology

This paper introduced Grad-CAM, a technique for producing visual explanations for decisions made by CNN-based models. It uses the gradients of the target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting that concept. This makes the model's decisions more transparent.

## Relevance to Primary Survey

Cited as the source of the idea of using gradient information as a selection criterion in the counterfactual contrastive learning model [88]. It shows how techniques from model interpretability can be repurposed to improve the training process itself.

# Reference [90]: Grounding Action Descriptions in Videos (Regneri et al., 2013)

## Key Concepts and Methodology

This paper introduced the TACOS dataset, which was created by taking the existing MPII-Compositive dataset (focused on cooking activities) and re-annotating it with natural language descriptions for various actions. This was done at multiple levels of detail using crowd-sourcing.

## Main Contributions and Findings

The TACOS dataset is notable for its domain specificity (culinary videos) and the relatively long average length of its videos (around 5 minutes).

## Limitations and Broader Context

The specific domain and long video length make it a particularly challenging dataset, and most NLVL methods report relatively low accuracy on it.

## Relevance to Primary Survey

Cited as the source of the TACOS dataset, one of the four commonly used benchmarks for evaluating NLVL models.

# Reference [91]: ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding (Heilbron et al., 2015)

## Key Concepts and Methodology

This paper introduced the ActivityNet dataset, a large-scale, open-domain benchmark for various human activity understanding tasks. The ActivityNet Captions dataset, which was derived from this, contains around 20,000 untrimmed videos from a wide variety of contexts and over 70,000 sentence queries.

## Limitations and Broader Context

The videos and the target moments within them have a large variance in length, making it a difficult and realistic dataset for evaluating the robustness of NLVL models.

## Relevance to Primary Survey

Cited as the source of the ActivityNet Captions dataset, which is the largest and most diverse benchmark currently used for NLVL, making it a crucial testbed for model generalization.

# Reference [92]: The New Data and New Challenges in Multimedia Research (Thomee et al., 2015)

## Key Concepts and Methodology

This paper discusses the landscape of new, large-scale multimedia datasets that were emerging at the time. It specifically details the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset, a massive collection of photos and videos.

## Relevance to Primary Survey

Cited as the source from which the videos for the DiDeMo dataset were collected. This highlights the data lineage of one of the key NLVL benchmarks.

# Reference [93]: Script Data for Attribute-Based Recognition of Composite Activities (Rohrbach et al., 2012)

## Key Concepts and Methodology

This paper introduced the MPII-Compositive dataset, which focuses on composite activities, primarily in the domain of cooking. It provides detailed script-like annotations for these activities.

## Relevance to Primary Survey

Cited as the original source dataset that was later re-annotated by Regneri et al. [90] to create the TACOS dataset for NLVL.

# Reference [94]: Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding (Sigurdsson et al., 2016)

## Key Concepts and Methodology

This paper introduced the Charades dataset, a unique dataset collected via crowdsourcing where people were asked to act out and record daily activities in their own homes based on scripts. It contains around 10,000 videos classified into 157 predefined activity categories.

## Relevance to Primary Survey

Cited as the original source dataset that was later re-labeled with natural language sentences by Gao et al. [28] to create the Charades-STA dataset, another standard benchmark for the NLVL task.

# Reference [95]: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks (Lu et al., 2019)

## Key Concepts and Methodology

This paper is a significant work in vision-language pre-training (VLP). It introduces ViLBERT, a model with two separate Transformer-based streams for processing visual and textual inputs, which are then fused through co-attentional transformer layers to learn joint representations.

## Relevance to Primary Survey

Cited in the survey's conclusion as a key example of powerful, recent work in vision-language understanding that had not yet been sufficiently attended to by NLVL researchers. The implication is that leveraging such pre-trained models could significantly boost performance.

# Reference [96]: LXMERT: Learning Cross-Modality Encoder Representations from Transformers (Tan & Bansal, 2019)

## Key Concepts and Methodology

This paper introduced LXMERT, another influential large-scale Transformer model for learning cross-modality representations. It features three encoders: a language encoder, an object-relationship encoder for the visual input, and a cross-modality encoder to fuse them. It was pre-trained on a large amount of image-text data.

## Relevance to Primary Survey

Cited in the conclusion as another important advancement in vision-language pre-training. The survey suggests that future NLVL models could benefit from adopting these powerful, pre-trained cross-modal encoders instead of learning them from scratch on smaller NLVL datasets.

# Reference [97]: VL-BERT: Pre-Training of Generic Visual-Linguistic Representations (Su et al., 2019)

## Key Concepts and Methodology

This paper presents VL-BERT, which adapts the successful BERT model to the vision-language domain. It uses a single-stream Transformer architecture where visual features (e.g., region proposals) and text embeddings are fed in as a single input sequence, allowing the model to learn deep, fused representations from the very first layer.

## Relevance to Primary Survey

Cited in the conclusion as part of a group of recent, powerful vision-language models whose techniques could benefit the NLVL field by providing much stronger feature representations.

# Reference [8]: Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training (Li et al., 2020)

## Key Concepts and Methodology

This paper proposed a universal encoder for vision and language, trained via cross-modal pre-training on three tasks: Masked Language Modeling (similar to BERT), Masked Object Classification, and Visual-Linguistic Matching. This multi-task pre-training helps it learn robust and generalizable representations.

## Relevance to Primary Survey

Cited in the conclusion to highlight a key direction for future work in NLVL, which is to move away from training models from scratch and instead fine-tune powerful, pre-trained visual-linguistic foundation models like this one.

# Reference [99]: Unified Vision-Language Pre-Training for Image Captioning and VQA (Zhou et al., 2020)

## Key Concepts and Methodology

This paper proposes a unified pre-training framework that is designed to be effective for both vision-language generation tasks (like image captioning) and understanding tasks (like visual question answering - VQA). It uses a shared multi-layer Transformer network for encoding and decoding.

## Relevance to Primary Survey

Cited in the survey's conclusion as a key example of recent work in VLP. The advancements from such models could directly address the limitation of existing NLVL models that often rely on outdated and separately trained feature extractors.

**Links of Referenced papers and Sources**

1. **Title:** Temporal Action Localization in Untrimmed Videos via Multi-Stage CNNs (Shou et al., 2016)
   * **Paper:** <https://arxiv.org/abs/1601.02129>
   * **GitHub:** Not available.
2. **Title:** Cascaded Boundary Regression for Temporal Action Detection (Gao et al., 2017)
   * **Paper:** <https://arxiv.org/abs/1705.01180>
   * **GitHub:** Not available.
3. **Title:** Rethinking the Faster R-CNN Architecture for Temporal Action Localization (Chao et al., 2018)
   * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2018/papers/Chao_Rethinking_the_Faster_CVPR_2018_paper.pdf>
   * **GitHub:** Not available.
4. **Title:** End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos (Buch et al., 2019)
   * **Paper:** <http://vision.stanford.edu/pdf/buch2017bmvc.pdf>
   * **GitHub:** Not available.
5. **Title:** Gaussian Temporal Awareness Networks for Action Localization (Long et al., 2019)
   * **Paper:** <https://arxiv.org/abs/1905.07413>
   * **GitHub:** [https://github.com/TencentYoutuResearch/ActionDetection-GTAN](https://www.google.com/search?q=https://github.com/TencentYoutuResearch/ActionDetection-GTAN&authuser=2)
6. **Title:** Visual to Text: Survey of Image and Video Captioning (Li et al., 2019)
   * **Paper:** <https://arxiv.org/abs/1904.06859>
   * **GitHub:** Not available.
7. **Title:** Reconstruction Network for Video Captioning (Wang et al., 2018)
   * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Reconstruction_Network_for_CVPR_2018_paper.pdf>
   * **GitHub:** Not available.
8. **Title:** Video Captioning with Transferred Semantic Attributes (Pan et al., 2017)
   * **Paper:** <https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/Video-Captioning-with-Transferred-Semantic-Attributes.pdf>
   * **GitHub:** Not available.
9. **Title:** Video Captioning with Attention-Based LSTM and Semantic Consistency (Gao et al., 2017)
   * **Paper:** <https://www.researchgate.net/publication/310820115_Attention-based_LSTM_with_Semantic_Consistency_for_Videos_Captioning>
   * **GitHub:** Not available.
10. **Title:** Video Captioning via Hierarchical Reinforcement Learning (Wang et al., 2018)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Video_Captioning_via_CVPR_2018_paper.pdf>
    * **GitHub:** [https://github.com/j-min/video-captioning-h-rl](https://www.google.com/search?q=https://github.com/j-min/video-captioning-h-rl&authuser=2)
11. **Title:** Weakly Supervised Dense Video Captioning (Shen et al., 2017)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2017/papers/Shen_Weakly_Supervised_Dense_CVPR_2017_paper.pdf>
    * **GitHub:** Not available.
12. **Title:** SST: Single-Stream Temporal Action Proposals (Buch et al., 2017)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2017/papers/Buch_SST_Single-Stream_Temporal_CVPR_2017_paper.pdf>
    * **GitHub:** [https://github.com/shyamal-b/SST](https://www.google.com/search?q=https://github.com/shyamal-b/SST&authuser=2)
13. **Title:** TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals (Gao et al., 2017)
    * **Paper:** <https://openaccess.thecvf.com/content_ICCV_2017/papers/Gao_TURN_TAP_Temporal_ICCV_2017_paper.pdf>
    * **GitHub:** <https://github.com/jiyanggao/TURN-TAP>
14. **Title:** DAPS: Deep Action Proposals for Action Understanding (Escorcia et al., 2016)
    * **Paper:** <https://link.springer.com/chapter/10.1007/978-3-319-46487-9_33>
    * **GitHub:** Not available.
15. **Title:** Play and Rewind: Context-Aware Video Temporal Action Proposals (Gao et al., 2020)
    * **Paper:** <https://openaccess.thecvf.com/content/WACV2022/papers/Hsieh_Contextual_Proposal_Network_for_Action_Localization_WACV_2022_paper.pdf>
    * **GitHub:** Not available.
16. **Title:** Learning Spatiotemporal Features with 3D Convolutional Networks (Tran et al., 2015)
    * **Paper:** <https://openaccess.thecvf.com/content_iccv_2015/papers/Tran_Learning_Spatiotemporal_Features_ICCV_2015_paper.pdf>
    * **GitHub:** Not available.
17. **Title:** Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (Carreira & Zisserman, 2017)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2017/papers/Carreira_Quo_Vadis_Action_CVPR_2017_paper.pdf>
    * **GitHub:** <https://github.com/deepmind/kinetics-i3d>
18. **Title:** Very Deep Convolutional Networks for Large-Scale Image Recognition (Simonyan & Zisserman, 2014)
    * **Paper:** <https://arxiv.org/abs/1409.1556>
    * **GitHub:** Not available.
19. **Title:** Fast R-CNN (Girshick, 2015)
    * **Paper:** <https://openaccess.thecvf.com/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf>
    * **GitHub:** <https://github.com/rbgirshick/fast-rcnn>
20. **Title:** Skip-Thought Vectors (Kiros et al., 2015)
    * **Paper:** <https://arxiv.org/abs/1506.06726>
    * **GitHub:** <https://github.com/ryankiros/skip-thoughts>
21. **Title:** Natural Language Processing with Python (Bird et al., 2009)
    * **Paper (Book):** <https://www.nltk.org/book/>
    * **GitHub:** <https://github.com/nltk/nltk>
22. **Title:** GloVe: Global Vectors for Word Representation (Pennington et al., 2014)
    * **Paper:** <https://aclanthology.org/D14-1162/>
    * **GitHub:** <https://github.com/stanfordnlp/GloVe>
23. **Title:** Long Short-Term Memory (Hochreiter & Schmidhuber, 1997)
    * **Paper:** <https://www.bioinf.jku.at/publications/older/2604.pdf>
    * **GitHub:** Not available.
24. **Title:** Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (Chung et al., 2014)
    * **Paper:** <https://arxiv.org/abs/1412.3555>
    * **GitHub:** Not available.
25. **Title:** Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos (Zhang et al., 2019)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3343031.3350927>
    * **GitHub:** [https://github.com/keep-learning-cpp/CMIN](https://www.google.com/search?q=https://github.com/keep-learning-cpp/CMIN&authuser=2)
26. **Title:** Localizing Moments in Video with Natural Language (Hendricks et al., 2017)
    * **Paper:** [https://openaccess.thecvf.com/content\_iccv\_2017/papers/Hendricks\_Localizing\_Moments\_in\_ICCV\_2017\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_iccv_2017/papers/Hendricks_Localizing_Moments_in_ICCV_2017_paper.pdf&authuser=2)
    * **GitHub:** [https://github.com/LisaAnne/MCN](https://www.google.com/search?q=https://github.com/LisaAnne/MCN&authuser=2)
27. **Title:** Localizing Moments in Video with Temporal Language (Hendricks et al., 2018)
    * **Paper:** <https://aclanthology.org/D18-1168/>
    * **GitHub:** [https://github.com/LisaAnne/Localizing-Moments-in-Video-with-Temporal-Language](https://www.google.com/search?q=https://github.com/LisaAnne/Localizing-Moments-in-Video-with-Temporal-Language&authuser=2)
28. **Title:** TALL: Temporal Activity Localization via Language Query (Gao et al., 2017)
    * **Paper:** [https://openaccess.thecvf.com/content\_iccv\_2017/papers/Gao\_TALL\_Temporal\_Activity\_ICCV\_2017\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_iccv_2017/papers/Gao_TALL_Temporal_Activity_ICCV_2017_paper.pdf&authuser=2)
    * **GitHub:** <https://github.com/jiyanggao/TALL>
29. **Title:** Attentive Moment Retrieval in Videos (Liu et al., 2018)
    * **Paper:** [https://dl.acm.org/doi/abs/10.1145/3209978.3210034](https://www.google.com/search?q=https://dl.acm.org/doi/abs/10.1145/3209978.3210034&authuser=2)
    * **GitHub:** Not available.
30. **Title:** Cross-Modal Moment Localization in Videos (Liu et al., 2018)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3240508.3240549>
    * **GitHub:** Not available.
31. **Title:** Multi-Modal Circulant Fusion for Video-to-Language and Backward (Wu & Han, 2018)
    * **Paper:** [https://openaccess.thecvf.com/content\_ECCV\_2018/papers/Ying\_Wu\_Multi-Modal\_Circulant\_Fusion\_ECCV\_2018\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_ECCV_2018/papers/Ying_Wu_Multi-Modal_Circulant_Fusion_ECCV_2018_paper.pdf&authuser=2)
    * **GitHub:** Not available.
32. **Title:** Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos (Liu et al., 2018)
    * **Paper:** <https://aclanthology.org/D18-1167/>
    * **GitHub:** Not available.
33. **Title:** Exploiting Temporal Relationships in Video Moment Localization with Natural Language (Zhang et al., 2019)
    * **Paper:** [https://openaccess.thecvf.com/content\_WACV\_2022/papers/Zhang\_Natural\_Language\_Video\_Moment\_Localization\_Through\_Query-Controlled\_Temporal\_Convolution\_WACV\_2022\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_WACV_2022/papers/Zhang_Natural_Language_Video_Moment_Localization_Through_Query-Controlled_Temporal_Convolution_WACV_2022_paper.pdf&authuser=2)
    * **GitHub:** Not available.
34. **Title:** MAC: Mining Activity Concepts for Language-Based Temporal Localization (Ge et al., 2019)
    * **Paper:** <https://arxiv.org/abs/1811.08925>
    * **GitHub:** [https://github.com/RunzhouGe/MAC-Net](https://www.google.com/search?q=https://github.com/RunzhouGe/MAC-Net&authuser=2)
35. **Title:** Semantic Proposal for Activity Localization in Videos via Sentence Query (Chen & Jiang, 2019)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3343031.3351042>
    * **GitHub:** Not available.
36. **Title:** Cross-Modal Video Moment Retrieval with Spatial and Language-Temporal Attention (Jiang et al., 2019)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3343031.3350905>
    * **GitHub:** Not available.
37. **Title:** Find and Focus: Retrieve and Localize Video Events with Natural Language Queries (Shao et al., 2018)
    * **Paper:** [https://openaccess.thecvf.com/content\_ECCV\_2018/papers/Dian\_Shao\_Find\_and\_Focus\_ECCV\_2018\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_ECCV_2018/papers/Dian_Shao_Find_and_Focus_ECCV_2018_paper.pdf&authuser=2)
    * **GitHub:** Not available.
38. **Title:** Multilevel Language and Vision Integration for Text-to-Clip Retrieval (Xu et al., 2019)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/5162>
    * **GitHub:** Not available.
39. **Title:** Interaction-Integrated Network for Natural Language Moment Localization (Ning et al., 2021)
    * **Paper:** <https://ieeexplore.ieee.org/document/9329705>
    * **GitHub:** Not available.
40. **Title:** Temporally Grounding Natural Sentence in Video (Chen et al., 2018)
    * **Paper:** <https://aclanthology.org/D18-1166/>
    * **GitHub:** Not available.
41. **Title:** Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction (Lin et al., 2020)
    * **Paper:** [https://ieeexplore.ieee.org/abstract/document/8962624/](https://www.google.com/search?q=https://ieeexplore.ieee.org/abstract/document/8962624/&authuser=2)
    * **GitHub:** Not available.
42. **Title:** DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization (Lu et al., 2019)
    * **Paper:** <https://aclanthology.org/D19-1518/>
    * **GitHub:** Not available.
43. **Title:** QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension (Yu et al., 2018)
    * **Paper:** <https://arxiv.org/abs/1804.09541>
    * **GitHub:** [https://github.com/hackiey/QANet-pytorch](https://www.google.com/search?q=https://github.com/hackiey/QANet-pytorch&authuser=2)
44. **Title:** Rethinking the Bottom-Up Framework for Query-Based Video Localization (Chen et al., 2020)
    * **Paper:** <https://arxiv.org/abs/2005.05234>
    * **GitHub:** Not available.
45. **Title:** Dense Regression Network for Video Grounding (Zeng et al., 2020)
    * **Paper:** <https://openaccess.thecvf.com/content_CVPR_2020/papers/Zeng_Dense_Regression_Network_for_Video_Grounding_CVPR_2020_paper.pdf>
    * **GitHub:** Not available.
46. **Title:** Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization (Liu et al., 2020)
    * **Paper:** <https://arxiv.org/abs/2008.01403>
    * **GitHub:** Not available.
47. **Title:** Fine-Grained Iterative Attention Network for Temporal Language Localization in Videos (Qu et al., 2020)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3394171.3413554>
    * **GitHub:** Not available.
48. **Title:** Attention on Attention for Image Captioning (Huang et al., 2019)
    * **Paper:** <https://openaccess.thecvf.com/content_ICCV_2019/papers/Huang_Attention_on_Attention_for_Image_Captioning_ICCV_2019_paper.pdf>
    * **GitHub:** <https://github.com/husthuaan/AoANet>
49. **Title:** MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment (Zhang et al., 2019)
    * **Paper:** <https://openaccess.thecvf.com/content_CVPR_2019/papers/Zhang_MAN_Moment_Alignment_Network_for_Natural_Language_Moment_Retrieval_via_CVPR_2019_paper.pdf>
    * **GitHub:** Not available.
50. **Title:** Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos (Yuan et al., 2019)
    * **Paper:** [https://proceedings.neurips.cc/paper/2019/file/b4a526523729c423f735625c158348df-Paper.pdf](https://www.google.com/search?q=https://proceedings.neurips.cc/paper/2019/file/b4a526523729c423f735625c158348df-Paper.pdf&authuser=2)
    * **GitHub:** Not available.
51. **Title:** TAN: Temporal Aggregation Network for Dense Multi-Label Action Recognition (Dai et al., 2019)
    * **Paper:** [https://openaccess.thecvf.com/content\_CVPR\_2020/papers/Dai\_TAN\_Temporal\_Aggregation\_Network\_for\_Dense\_Multi-Label\_Action\_Recognition\_CVPR\_2020\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2020/papers/Dai_TAN_Temporal_Aggregation_Network_for_Dense_Multi-Label_Action_Recognition_CVPR_2020_paper.pdf&authuser=2)
    * **GitHub:** Not available.
52. **Title:** Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language (Zhang et al., 2020)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/6984>
    * **GitHub:** [https://github.com/Panda-Peter/2D-TAN](https://www.google.com/search?q=https://github.com/Panda-Peter/2D-TAN&authuser=2)
53. **Title:** Dual Path Interaction Network for Video Moment Localization (Wang et al., 2020)
    * **Paper:** <https://dl.acm.org/doi/abs/10.1145/3394171.3413536>
    * **GitHub:** Not available.
54. **Title:** To Find Where You Talk: Temporal Sentence Localization in Video with Attention Based Location Regression (Yuan et al., 2019)
    * **Paper:** [https://openaccess.thecvf.com/content\_CVPR\_2019/papers/Yuan\_To\_Find\_Where\_You\_Talk\_Temporal\_Sentence\_Localization\_in\_Video\_CVPR\_2019\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2019/papers/Yuan_To_Find_Where_You_Talk_Temporal_Sentence_Localization_in_Video_CVPR_2019_paper.pdf&authuser=2)
    * **GitHub:** Not available.
55. **Title:** Local-Global Video-Text Interactions for Temporal Grounding (Mun et al., 2020)
    * **Paper:** <https://openaccess.thecvf.com/content_CVPR_2020/papers/Mun_Local-Global_Video-Text_Interactions_for_Temporal_Grounding_CVPR_2020_paper.pdf>
    * **GitHub:** Not available.
56. **Title:** BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2018)
    * **Paper:** <https://aclanthology.org/N19-1423/>
    * **GitHub:** <https://github.com/google-research/bert>
57. **Title:** Hadamard Product for Low-Rank Bilinear Pooling (Kim et al., 2016)
    * **Paper:** <https://arxiv.org/abs/1610.04325>
    * **GitHub:** Not available.
58. **Title:** Non-Local Neural Networks (Wang et al., 2018)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Non-Local_Neural_Networks_CVPR_2018_paper.pdf>
    * **GitHub:** <https://github.com/facebookresearch/video-nonlocal-net>
59. **Title:** Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos (Chen et al., 2020)
    * **Paper:** <https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123490324.pdf>
    * **GitHub:** Not available.
60. **Title:** Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language (Chen & Jiang, 2020)
    * **Paper:** <https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123650596.pdf>
    * **GitHub:** <https://github.com/forwchen/HVTG>
61. **Title:** Ring Loss: Convex Feature Normalization for Face Recognition (Zheng et al., 2018)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2018/papers/Zheng_Ring_Loss_Convex_CVPR_2018_paper.pdf>
    * **GitHub:** [https://github.com/ydwen/Ring-Loss](https://www.google.com/search?q=https://github.com/ydwen/Ring-Loss&authuser=2)
62. **Title:** Attention-Aware Deep Reinforcement Learning for Video Face Recognition (Rao et al., 2017)
    * **Paper:** <https://openaccess.thecvf.com/content_ICCV_2017/papers/Rao_Attention-Aware_Deep_Reinforcement_ICCV_2017_paper.pdf>
    * **GitHub:** Not available.
63. **Title:** Reinforcement Cutting-Agent Learning for Video Object Segmentation (Han et al., 2018)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2018/papers/Han_Reinforcement_Cutting-Agent_Learning_CVPR_2018_paper.pdf>
    * **GitHub:** Not available.
64. **Title:** Deep Reinforcement Learning for General Video Game AI (Torrado et al., 2018)
    * **Paper:** <https://nyuscholars.nyu.edu/en/publications/deep-reinforcement-learning-for-general-video-game-ai>
    * **GitHub:** [https://github.com/rubenrt/gvgai-gym](https://www.google.com/search?q=https://github.com/rubenrt/gvgai-gym&authuser=2)
65. **Title:** Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos (He et al., 2019)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/4859>
    * **GitHub:** [https://github.com/lhew/RWM](https://www.google.com/search?q=https://github.com/lhew/RWM&authuser=2)
66. **Title:** Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model (Wang et al., 2019)
    * **Paper:** <https://openaccess.thecvf.com/content_CVPR_2019/papers/Wang_Language-Driven_Temporal_Activity_Localization_A_Semantic_Matching_Reinforcement_Learning_Model_CVPR_2019_paper.pdf>
    * **GitHub:** Not available.
67. **Title:** Tree-Structured Policy Based Progressive Reinforcement Learning for Temporally Language Grounding in Video (Wu et al., 2020)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/6959>
    * **GitHub:** [https://github.com/wubiaoz/TSP-PRL](https://www.google.com/search?q=https://github.com/wubiaoz/TSP-PRL&authuser=2)
68. **Title:** Tripping Through Time: Efficient Localization of Activities in Videos (Hahn et al., 2020)
    * **Paper:** [https://openaccess.thecvf.com/content\_WACV\_2020/papers/Hahn\_Tripping\_Through\_Time\_Efficient\_Localization\_of\_Activities\_in\_Videos\_WACV\_2020\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_WACV_2020/papers/Hahn_Tripping_Through_Time_Efficient_Localization_of_Activities_in_Videos_WACV_2020_paper.pdf&authuser=2)
    * **GitHub:** Not available.
69. **Title:** Gated-Attention Readers for Text Comprehension (Dhingra et al., 2016)
    * **Paper:** <https://aclanthology.org/P17-1018/>
    * **GitHub:** <https://github.com/bdhingra/ga-reader>
70. **Title:** Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization (Cao et al., 2020)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3394171.3413550>
    * **GitHub:** Not available.
71. **Title:** STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization (Cao et al., 2020)
    * **Paper:** <https://dl.acm.org/doi/abs/10.1145/3394171.3413721>
    * **GitHub:** Not available.
72. **Title:** Localizing Natural Language in Videos (Chen et al., 2019)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/4836>
    * **GitHub:** Not available.
73. **Title:** ExCL: Extractive Clip Localization Using Natural Language Descriptions (Ghosh et al., 2019)
    * **Paper:** <https://aclanthology.org/N19-1198/>
    * **GitHub:** [https://github.com/SohamGhosh/ExCL](https://www.google.com/search?q=https://github.com/SohamGhosh/ExCL&authuser=2)
74. **Title:** Temporally Grounding Language Queries in Videos by Contextual Boundary-Aware Prediction (Wang et al., 2020)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/6890>
    * **GitHub:** [https://github.com/wangkai-cuhk/CBP](https://www.google.com/search?q=https://github.com/wangkai-cuhk/CBP&authuser=2)
75. **Title:** Proposal-Free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention (Rodriguez-Opazo et al., 2020)
    * **Paper:** [https://openaccess.thecvf.com/content\_WACV\_2020/papers/Rodriguez-Opazo\_Proposal-Free\_Temporal\_Moment\_Localization\_of\_a\_Natural-Language\_Query\_in\_Video\_WACV\_2020\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_WACV_2020/papers/Rodriguez-Opazo_Proposal-Free_Temporal_Moment_Localization_of_a_Natural-Language_Query_in_Video_WACV_2020_paper.pdf&authuser=2)
    * **GitHub:** Not available.
76. **Title:** Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models (Hershey & Olsen, 2007)
    * **Paper:** [https://www.merl.com/publications/docs/TR2007-063.pdf](https://www.google.com/search?q=https://www.merl.com/publications/docs/TR2007-063.pdf&authuser=2)
    * **GitHub:** Not available.
77. **Title:** DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video (Rodriguez-Opazo et al., 2021)
    * **Paper:** [https://openaccess.thecvf.com/content/CVPR2021/papers/Rodriguez-Opazo\_DORi\_Discovering\_Object\_Relationships\_for\_Moment\_Localization\_of\_a\_Natural\_CVPR\_2021\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content/CVPR2021/papers/Rodriguez-Opazo_DORi_Discovering_Object_Relationships_for_Moment_Localization_of_a_Natural_CVPR_2021_paper.pdf&authuser=2)
    * **GitHub:** Not available.
78. **Title:** Span-Based Localizing Network for Natural Language Video Localization (Zhang et al., 2020)
    * **Paper:** <https://aclanthology.org/2020.acl-main.448/>
    * **GitHub:** Not available.
79. **Title:** Weakly Supervised Dense Event Captioning in Videos (Duan et al., 2018)
    * **Paper:** [https://openaccess.thecvf.com/content\_cvpr\_2018/papers/Duan\_Weakly\_Supervised\_Dense\_CVPR\_2018\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_cvpr_2018/papers/Duan_Weakly_Supervised_Dense_CVPR_2018_paper.pdf&authuser=2)
    * **GitHub:** Not available.
80. **Title:** Iterative Approximation of Fixed Points of Lipschitzian Strictly Pseudocontractive Mappings (Chidume, 1987)
    * **Paper:** [https://www.ams.org/proc/1987-099-02/S0002-9939-1987-0869344-3/S0002-9939-1987-0869344-3.pdf](https://www.google.com/search?q=https://www.ams.org/proc/1987-099-02/S0002-9939-1987-0869344-3/S0002-9939-1987-0869344-3.pdf&authuser=2)
    * **GitHub:** Not available.
81. **Title:** Weakly Supervised Video Moment Retrieval from Text Queries (Mithun et al., 2019)
    * **Paper:** <https://openaccess.thecvf.com/content_CVPR_2019/papers/Mithun_Weakly_Supervised_Video_Moment_Retrieval_From_Text_Queries_CVPR_2019_paper.pdf>
    * **GitHub:** Not available.
82. **Title:** WSLLN: Weakly Supervised Natural Language Localization Networks (Gao et al., 2019)
    * **Paper:** <https://aclanthology.org/D19-1520/>
    * **GitHub:** Not available.
83. **Title:** Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos (Wu et al., 2020)
    * **Paper:** <https://dl.acm.org/doi/10.1145/3394171.3413551>
    * **GitHub:** Not available.
84. **Title:** Reinforcement Learning: An Introduction (Sutton & Barto, 2018)
    * **Paper (Book):** <http://incompleteideas.net/book/the-book-2nd.html>
    * **GitHub:** <https://github.com/ShangtongZhang/reinforcement-learning-an-introduction>
85. **Title:** Weakly-Supervised Video Moment Retrieval via Semantic Completion Network (Lin et al., 2020)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/6789>
    * **GitHub:** Not available.
86. **Title:** Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos (Zhang et al., 2020)
    * **Paper:** [https://openaccess.thecvf.com/content\_CVPR\_2020/papers/Zhang\_Regularized\_Two-Branch\_Proposal\_Networks\_for\_Weakly-Supervised\_Moment\_Retrieval\_in\_CVPR\_2020\_paper.pdf](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2020/papers/Zhang_Regularized_Two-Branch_Proposal_Networks_for_Weakly-Supervised_Moment_Retrieval_in_CVPR_2020_paper.pdf&authuser=2)
    * **GitHub:** Not available.
87. **Title:** VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval (Ma et al., 2020)
    * **Paper:** <https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123690460.pdf>
    * **GitHub:** Not available.
88. **Title:** Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding (Zhang et al., 2020)
    * **Paper:** <https://proceedings.neurips.cc/paper/2020/file/d27b95cac4c27feb850aaa4070cc4675-Paper.pdf>
    * **GitHub:** Not available.
89. **Title:** Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization (Selvaraju et al., 2017)
    * **Paper:** <https://openaccess.thecvf.com/content_iccv_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf>
    * **GitHub:** <https://github.com/ramprs/grad-cam>
90. **Title:** Grounding Action Descriptions in Videos (Regneri et al., 2013)
    * **Paper:** <https://aclanthology.org/Q13-1003/>
    * **GitHub:** Not available.
91. **Title:** ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding (Heilbron et al., 2015)
    * **Paper:** <https://openaccess.thecvf.com/content_cvpr_2015/papers/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.pdf>
    * **GitHub:** Not available.
92. **Title:** The New Data and New Challenges in Multimedia Research (Thomee et al., 2015)
    * **Paper:** <https://arxiv.org/abs/1503.01817>
    * **GitHub:** Not available.
93. **Title:** Script Data for Attribute-Based Recognition of Composite Activities (Rohrbach et al., 2012)
    * **Paper:** <https://link.springer.com/chapter/10.1007/978-3-642-33718-5_14>
    * **GitHub:** Not available.
94. **Title:** Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding (Sigurdsson et al., 2016)
    * **Paper:** <https://arxiv.org/abs/1604.01753>
    * **GitHub:** Not available.
95. **Title:** ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks (Lu et al., 2019)
    * **Paper:** <https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf>
    * **GitHub:** <https://github.com/facebookresearch/vilbert-multi-task>
96. **Title:** LXMERT: Learning Cross-Modality Encoder Representations from Transformers (Tan & Bansal, 2019)
    * **Paper:** <https://aclanthology.org/D19-1514/>
    * **GitHub:** <https://github.com/airsplay/lxmert>
97. **Title:** VL-BERT: Pre-Training of Generic Visual-Linguistic Representations (Su et al., 2019)
    * **Paper:** <https://arxiv.org/abs/1908.08530>
    * **GitHub:** <https://github.com/jackroos/VL-BERT>
98. **Title:** Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training (Li et al., 2020)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/6770>
    * **GitHub:** <https://github.com/microsoft/Unicoder>
99. **Title:** Unified Vision-Language Pre-Training for Image Captioning and VQA (Zhou et al., 2020)
    * **Paper:** <https://ojs.aaai.org/index.php/AAAI/article/view/6962>
    * **GitHub:** <https://github.com/LuoweiZhou/VLP>