# Technical Summary of Joint Event Detection and Description in Continuous Video Streams (JEDDi-Net)

Part I: Problem Formulation, Methods, and Evidence

Kai-Yu Lu

2025/08/30

## 1 Research Problem and Motivation

Dense video captioning in continuous streams requires two coupled capabilities: detecting temporally localized events and generating a natural language description for each detected event. Earlier pipelines handled temporal proposal generation and sentence decoding with separate modules, which prevented captioning signals from refining proposal boundaries and often summarized proposal content with coarse, shared states. The work summarized here introduces an end-to-end trainable architecture, JEDDi-Net, that jointly detects temporal events and decodes their captions while modeling both visual context and language history across events.

**Paper facts** **Title:** Joint Event Detection and Description in Continuous Video Streams.
**Authors:** Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, Kate Saenko.

## 2 Related Work

Temporal activity detection evolved from trimmed classification to localizing start and end times in untrimmed videos using sliding windows, recurrent encoders, and proposal networks. Video captioning progressed from template filling to CNN–RNN models with attention and hierarchical language models for paragraph generation. Dense video captioning on ActivityNet Captions frames joint temporal localization and sentence generation; prior methods typically decoupled these stages and represented proposals using shared RNN states. JEDDi-Net integrates an R-C3D style proposal module with a hierarchical captioner that conditions on both visual context and prior sentences.

**Terminology** *C3D* refers to a 3D convolutional network encoding spatiotemporal features. *SPN* denotes a Segment Proposal Network that classifies temporal anchors as activity or background and regresses segment boundaries. *tIoU* means Temporal Intersection-over-Union used to match proposals to ground-truth segments. *SoI pooling* indicates 3D Segment-of-Interest pooling that extracts fixed-size features from variable-length temporal segments. *Controller LSTM* and *Captioner LSTM* are the high-level and low-level recurrent decoders that encode cross-event context and generate words, respectively.

# 3 Dataset Construction

### ActivityNet Captions

Roughly twenty thousand untrimmed videos with train, validation, and test splits near 50%, 25%, and 25%. Each video contains multiple sentence-annotated temporal segments. Vocabulary retains tokens with frequency at least five. Frames are resized to $112 \times 112$, and fixed-length buffers of $L = 768$ frames are sampled at 3 fps with zero-padding if needed. Maximum caption length $K = 30$ covers the vast majority of sentences. Anchor scales are selected by cross-validation.

### TACoS-MultiLevel

Cooking videos with dense temporal descriptions. A common split uses 143 training and 42 test videos. All words are kept for the vocabulary. Maximum caption length is 15. Frames are sampled at 5 fps.

# 4 Query Protocol and Task Definitions

**Task:** Given a continuous video $V$, the system outputs a set of temporal segments paired with a sentence for each segment, forming a coherent multi-sentence description aligned to time. **Evaluation:** Detection is measured by area under the Average Recall versus Average Number of proposals curve (AUC) across tIoU thresholds. Captioning is scored by BLEU-1 to BLEU-4, METEOR, CIDEr, and ROUGE-L, computed on proposals that meet tIoU alignment thresholds with ground truth.

# 5 Modeling Approach

## 5.1 Video Feature Encoding

The input buffer of $L$ RGB frames with height $H$ and width $W$ is encoded by a fully convolutional C3D backbone, producing feature maps $C_{\text{conv5b}} \in \mathbb{R}^{512 \times L/8 \times H/16 \times W/16}$. Global max pooling followed by a fully connected layer (`fc6`) yields a video-level context vector $I_c$. Variable-length proposal features $I_{p,t}$ are extracted by 3D SoI pooling over $C_{\text{conv5b}}$ and then projected by `fc6`.

## 5.2 Segment Proposal Network

On top of $C_{\text{conv5b}}$, temporal features are formed by additional 3D convolutions and temporal pooling, yielding a sequence $C_{\text{tpn}}$. For each temporal location and each anchor with center $c_i$ and length $l_i$, the SPN predicts an activity label and offsets to the target center and length. Anchors with tIoU at least 0.7 to a ground-truth segment, or the best match per segment, are positives; anchors with tIoU at most 0.3 are negatives. Mini-batches are balanced 1:1. At inference time, non-maximum suppression with tIoU 0.7 refines proposals.

## 5.3 Hierarchical Captioning

A controller LSTM summarizes the video context $I_c$ and the previous sentence representation $S_{p,t-1}$ into a topic vector $h_t^c$. A two-layer captioner LSTM decodes tokens for the current proposal using $(I_{p,t}, h_t^c)$ and the embedded word history. $S_{p,t-1}$ is obtained by mean-pooling the word embeddings of the previous decoded sentence. Proposals are decoded in ascending end-time order to preserve narrative flow.

### 5.4 End-to-End Optimization

The captioner is first pre-trained on ground-truth proposal features extracted from a separately trained SPN to avoid low-diversity batches. The entire network is then fine-tuned end-to-end with a reduced learning rate so that captioning gradients improve proposal features and boundaries through the shared backbone.

### 5.5 Mathematical Formulation and Explanations

**Anchor-to-target transformation.**

$$\delta c_i^* = \frac{c_i^* - c_i}{l_i}, \quad \delta l_i^* = \log\left(\frac{l_i^*}{l_i}\right) \tag{1}$$

*Symbols:* $(c_i, l_i)$ are center and length of the $i$-th anchor; $(c_i^*, l_i^*)$ are the matched ground-truth parameters; $(\delta c_i^*, \delta l_i^*)$ are the target offsets. *Background:* one-dimensional adaptation of bounding-box regression. *Intuition:* centers are normalized by anchor length and lengths are regressed in log-space for scale stability. *Role:* defines regression targets for temporal boundary refinement.

**Smooth L1 regression loss.**

$$L_{\text{reg}}(x) = \mathbf{1}(|x| < 1) \cdot \tfrac{1}{2}x^2 + \mathbf{1}(|x| \geq 1) \cdot (|x| - \tfrac{1}{2}) \tag{2}$$

*Symbols:* $\mathbf{1}(\cdot)$ is the indicator function. *Background:* robust loss from Fast R-CNN. *Intuition:* quadratic near zero for precision and linear for outliers. *Role:* penalizes center and length offset errors.

**SPN joint loss over a mini-batch.**

$$L_{\text{spn}} = \frac{1}{M}\sum_{i=1}^{M}\Big[L_{\text{cls}}(\hat{a}_i, a_i^*) + a_i^*\big(L_{\text{reg}}(\hat{\delta c}_i - \delta c_i^*) + L_{\text{reg}}(\hat{\delta l}_i - \delta l_i^*)\big)\Big] \tag{3}$$

*Symbols:* $M$ is batch size; $\hat{a}_i$ predicted activity probability; $a_i^* \in \{0, 1\}$ ground-truth label; $\hat{\delta c}_i, \hat{\delta l}_i$ predicted offsets. *Background:* multitask classification and regression. *Intuition:* boundaries are regressed only for positive anchors. *Role:* trains the SPN to propose accurate temporal windows.

**Controller LSTM.**

$$\begin{bmatrix} f_t^c \\ i_t^c \\ o_t^c \end{bmatrix} = \sigma\left(\begin{bmatrix} W_f^c \\ W_i^c \\ W_o^c \end{bmatrix}\begin{bmatrix} S_{p,t-1} \\ I_c \\ h_{t-1}^c \end{bmatrix} + \begin{bmatrix} b_f^c \\ b_i^c \\ b_o^c \end{bmatrix}\right) \tag{4}$$

$$\tilde{c}_t^c = \tanh\Big(W_c^c[S_{p,t-1}, I_c, h_{t-1}^c] + b_c^c\Big) \tag{5}$$

$$c_t^c = i_t^c \odot \tilde{c}_t^c + f_t^c \odot c_{t-1}^c \tag{6}$$

$$h_t^c = o_t^c \odot \tanh(c_t^c) \tag{7}$$

*Symbols:* $S_{p,t-1}$ mean-pooled embedding of the previous sentence; $I_c$ video context; $h_t^c, c_t^c$ hidden and cell states; $\sigma$ logistic function; $\odot$ elementwise product. *Intuition:* fuses prior language and global visual cues into a topic vector $h_t^c$ for guiding the next sentence. *Role:* provides cross-event context to the captioner.

**Captioner LSTM second layer with fusion.**

$$\begin{bmatrix} f_k \\ i_k \\ o_k \end{bmatrix} = \sigma\left( \begin{bmatrix} W_f \\ W_i \\ W_o \end{bmatrix} \begin{bmatrix} h_k^{(1)} \\ I_{p,t} \\ h_t^c \\ h_{k-1}^{(2)} \end{bmatrix} + \begin{bmatrix} b_f \\ b_i \\ b_o \end{bmatrix} \right) \tag{8}$$

$$\tilde{c}_k = \tanh\Big(W_c[h_k^{(1)}, I_{p,t}, h_t^c, h_{k-1}^{(2)}] + b_c\Big) \tag{9}$$

$$c_k = i_k \odot \tilde{c}_k + f_k \odot c_{k-1} \tag{10}$$

$$h_k^{(2)} = o_k \odot \tanh(c_k) \tag{11}$$

*Symbols:* $h_k^{(1)}$ is the first-layer hidden state at word $k$; $I_{p,t}$ proposal feature; $h_k^{(2)}$ second-layer hidden state. *Intuition:* integrates local proposal evidence and global context at every decoding step. *Role:* produces logits for next-word prediction.

**Captioning objective.**

$$L_{\text{caption}} = -\frac{1}{KT} \sum_{t,k} \log P\big(w_k^t \mid I_{p,t},\, h_t^c,\, w_1^t, \ldots, w_{k-1}^t\big) \tag{12}$$

*Symbols:* $w_k^t$ ground-truth token at position $k$ of sentence $t$; $K$ maximum sentence length; $T$ number of proposals. *Background:* cross-entropy for sequence modeling. *Intuition:* maximizes likelihood of reference words given fused context. *Role:* training objective for the captioning decoder.

**Total loss.**

$$L_{\text{total}} = L_{\text{spn}} + \lambda\, L_{\text{caption}}, \quad \lambda = 1 \tag{13}$$

*Symbols:* $\lambda$ balances detection and captioning losses. *Role:* end-to-end objective that couples detection and captioning through shared features.

# 6 Empirical Results

## 6.1 Evaluation Protocols

For proposals, the AUC of Average Recall versus Average Number of proposals is reported at fixed tIoU values and averaged across thresholds from 0.5 to 0.95. For captions, BLEU-1 through BLEU-4, METEOR, CIDEr, and ROUGE-L are averaged across tIoU thresholds such as 0.3, 0.5, 0.7, and 0.9 using the top 1000 proposals per video.

## 6.2 ActivityNet Captions

**Proposal detection:** A separately trained SPN attains strong AUC; after joint training with the captioner, both the high-tIoU AUC and the average AUC across thresholds improve, indicating better boundary regression driven by language gradients.

**Dense captioning:** Compared with prior context models, JEDDi-Net improves METEOR and CIDEr while remaining competitive in BLEU-4. Ablations show consistent gains from joint training and further gains

from the hierarchical context that conditions on both $I_c$ and $S_{p,t-1}$.

**tIoU sensitivity:** As tIoU increases to moderate values, BLEU, METEOR, and ROUGE-L improve due to better alignment; very high tIoU reduces candidate proposals and can lower n-gram overlap, while CIDEr often increases as surviving captions are more distinctive.

### 6.3 TACoS-MultiLevel

**Proposals:** Average AUC improves after joint training.

**Captions:** The model reports the first dense captioning results on this benchmark with hierarchical context yielding higher BLEU-4, METEOR, CIDEr, and ROUGE-L than separate-training baselines. A trimmed captioning upper bound on the same annotations shows higher BLEU-4, underscoring the inherent difficulty of dense captioning with temporal localization.

### 6.4 Qualitative Analysis

Generated narratives respect temporal ordering and maintain cross-sentence coherence through the controller LSTM. Occasional lexical substitutions such as synonyms can depress BLEU-4 despite semantic adequacy, which highlights metric sensitivity to exact n-gram overlap.

## 7 Summary

### Innovation versus prior methods

- End-to-end joint optimization couples proposal quality to captioning loss rather than treating them as separate stages.
- Proposal features are obtained by 3D SoI pooling from shared C3D maps, avoiding coarse shared states across proposals.
- A hierarchical captioner fuses visual context and language history via a controller LSTM, promoting coherent multi-sentence output.

### Limitations and Future Directions

**Metric sensitivity:** Heavy reliance on n-gram overlap can undervalue semantically correct paraphrases.

**Decoding strategy:** Greedy decoding is used; beam search and coverage mechanisms may improve long-range fluency.

**Object grounding:** Misses on small or fast objects suggest that tighter object–language alignment or spatiotemporal transformers could help.

**Future:** Integrate beam search, incorporate object detectors or transformer backbones, and explore reinforcement learning to optimize sequence-level objectives directly.

## Technical Details (Reproducibility)

**Backbone and inputs:** C3D with $L = 768$, $H = W = 112$, 3 fps for ActivityNet and 5 fps for TACoS; anchors with multiple temporal scales.

**Training:** SPN initialized from a large-scale pretraining of C3D; captioner pre-trained on ground-truth proposal features; joint fine-tuning with $L_{\text{total}}$ and non-maximum suppression at tIoU 0.7.

**Evaluation:** Top-1000 proposals per video; metrics averaged across multiple tIoUs.