映像からのストーリー生成: イベント選択器と文生成器の同時学習

西村太一 [1]　橋本敦史 [2]　牛久祥孝 [2]　森信介 [3]
[1] 京都大学大学院 情報学情報学研究科 [2] オムロンサイニックエックス株式会社
[3] 京都大学 学術情報メディアセンター
taichitary@gmail.com　{atsushi.hashimoto,yoshitaka.ushiku}@sinicx.com
forest@i.kyoto-u.ac.jp

概要

本研究では，映像からストーリーを生成すること，つまり，映像から記述すべき重要なイベントを過不足なく検出し，その説明文を生成することを目的とする．この課題は，Dense Video Captioning (DVC) として取り組まれてきたが，既存モデルは主に映像から密にイベントを予測することに焦点を当てており，ストーリー性について考慮したものではなかった．そこで本研究では，映像からストーリーを生成する上で従来の方法と異なるアプローチを取る．それは，既存の DVC モデルの密な出力をイベント候補の集合とみなし，この集合から適切なイベントを選択しつつ文を生成する手法である．実験では，提案手法と従来モデルを比較して優れた性能を示しただけでなく，正しくストーリーを生成できていることも分かった．

1　はじめに

数秒から数十秒の映像からその説明文を生成することを目的とする Video Captioning は，近年急速に発展を遂げてきた [1, 2, 3]．一方で，比較的長い映像をもとに，映像からストーリーを生成すること，つまり映像から記述すべき重要なイベントを検出し，その説明文を生成すること，は未だ発展途上の挑戦的な課題である．この技術を確立できれば，自然言語をクエリとした動画のイベントの検索も可能となるし，出力したイベントと文の組み合わせは動画の概要を把握する上でも有用であろう．本研究では，映像からストーリーを生成する課題に取り組む．

この課題は，Vision and Language 分野において Dense Video Captioning [4] (DVC) として定式化され取り組まれており，近年は Transformer [5] をベースにしたモデルが提案されている [6, 7]．しかし，こうしたモデルは入力の映像について密にイベントを予測することに焦点を当て取り組まれており，ストーリー性について考慮したものではなかった．Fujuta ら [8] によると，100 を超えるような過剰なイベントの出力を行うことが報告されている．これは検索の観点からは有用であるかもしれないが，人が読んで理解するようなストーリー性の観点からは不適切であろう．

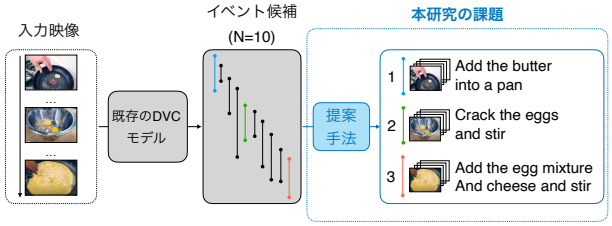


図 1　本研究の課題の概要．既存の DVC モデルが出力したイベント候補集合からストーリーを構成する上で欠かせないイベントを選択しつつ文を生成する．

そこで本研究では，映像からストーリーを生成するために，従来の方法と異なるアプローチを取る．それは，既存の DVC モデルの出力を**イベント候補の集合**と見なし，この集合から適切なイベント列を抽出するという方法である (図 1)．節 2 ではこのアプローチがストーリー生成に有効であることを既存の DVC モデルの出力をもとに議論し，この分析をもとに，節 3 にてイベントを候補から選択するイベント選択器と選択したイベントから文を生成する文生成器を同時学習する手法を提案する．

実験では，DVC においてよく利用されるデータセットの 1 つである YouCook2 [9] を用いて提案手法を評価した．その結果，従来の DVC モデルと比較して正しくストーリーを生成できていることが分かった．

2 既存の DVC モデルの出力の分析

既存の DVC モデルは正解のイベント数を大幅に超過してイベントを出力することが Fujita ら [8] によって指摘されている．そこで，我々は出力のイベント集合から正しいイベントを選択できれば正しいストーリーが得られるのではないかと考えた．これを検証するため，映像中の正解のイベントに対して，DVC モデルの出力集合の中で最もイベント間の重なり，つまり temporal Intersection over Union (tIoU) が大きいものを選択した時の評価結果を調査する．本研究ではこの手法を**オラクル**と呼ぶこととする．なお，データセットとしては YouCook2 [9] を用いて，DVC モデルとして PDVC [7] を検証に用いる．PDVC を選んだのは，現状このモデルが DVC において最も性能が高く，イベント候補数 N をあらかじめ設定できるからである[1)]．

評価尺度． 評価には，DVC の評価によく利用される 2 種類の評価尺度を用いる．1 つ目は DVC そのものと共に提案されたもので，本研究では dvc_eval [4] と呼ぶ．この評価尺度では，予測したイベント集合と正解のイベント集合の全ての組み合わせの tIoU を計算し，それが閾値 $\theta \in \{0.3, 0.5, 0.7, 0.9\}$ を超えた場合に METEOR [10] や CIDEr-D [11] などの文の自動評価尺度を計算し，その平均を取ることで評価される．この評価尺度は個々の予測イベントに対するイベント自体や生成文の正しさは評価できるが，動画全体でのストーリー性を考慮する仕組みにはなっていない．これを解決する評価尺度として，我々は SODA [8] も評価する．この評価尺度では，正解のイベントと予測イベントの間で動画全体のストーリー性を考慮したスコア (tIoU と文の自動評価尺度の積) が最大となる対応関係を動的計画法を用いて探索し評価を行う．動画全体のストーリー性を評価するのに適しているため，本研究では dvc_eval より SODA を重視する．

2.1 評価結果

定量的評価． 表 1 にオラクルの評価結果を，表 2 にイベントの予測結果を示す．PDVC と比較して，両方の評価尺度で高い性能を示している．イベント予測結果において強調すべき点として，閾値が 0.7 から 0.9 にかけてオラクルの評価結果が大きく下がることである．つまり，正解のイベントに対してオラクルの選択をしても，tIoU が 0.9 を超えるものは多くないことが分かる．

表 1 オラクルと PDVC の DVC 評価結果．

	dvc_eval			SODA	
	BLEU	METEOR	CIDEr-D	METEOR	CIDEr-D
PDVC	0.89	4.52	21.50	3.98	25.30
オラクル					
N=25	0.58	6.09	27.12	7.62	26.32
N=50	0.84	6.92	31.63	8.83	29.93
N=100	0.97	7.68	36.26	9.64	35.08
N=200	**1.10**	**8.15**	**38.60**	**10.43**	**36.89**

表 2 オラクルと PDVC のイベント予測結果．

	Recall				Precision			
	0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
PDVC	46.8	29.7	14.9	2.1	72.4	40.2	16.0	1.9
オラクル								
N=25	86.4	65.7	31.6	3.6	90.7	67.7	31.9	3.6
N=50	92.8	79.0	47.2	6.0	94.8	79.8	47.4	6.0
N=100	98.2	90.3	60.8	8.0	98.2	90.3	60.8	8.0
N=200	**99.3**	**95.9**	**76.2**	**12.6**	**99.5**	**95.9**	**76.2**	**12.6**

オラクルのイベントの tIoU の分布． 図 2 にオラクルと正解のイベント間での tIoU の分布を示す．学習データ，検証データともに N を十分に増やすと，オラクルの選択するイベントの tIoU は大きくなることが分かる．例えば，$N = 100$ において，オラクルの選択するイベントの tIoU の平均はそれぞれ 74.6%，70.4%であり，多くのイベントが 0.7 から 0.9 の間にあることが分かる．

3 提案手法

前節の分析において，イベント候補の中には正解のイベントと高い tIoU を持つイベントが存在し，それを選ぶことができれば正しいストーリーが得られることが分かった．これに基づき，本節では，イベントを選択するイベント選択器とそれに対応する文を生成する文生成器を同時に学習しストーリーを生成する手法を提案する．図 3 に提案手法の概要を示す．入力としてイベント候補列 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_N)$ が与えられる．期待する出力は，$\boldsymbol{X}$ 中のオラクルで選ぶイベントの添字とそれに対応する文の組 $(\boldsymbol{C}, \boldsymbol{Y}) = ((c_1, \boldsymbol{y}_1), \ldots, (c_t, \boldsymbol{y}_t), \ldots, (c_T, \boldsymbol{y}_T))$ である (T はイベント数)．提案手法は，t 番目においてイベントを 1 つ選択し，それに対応する文を生成する．この処理をイベント選択器がイベントの出力を終了するまで再帰的に繰り返す．

1) PDVC では，イベント候補数 N を設定してモデルを学習し，予測時はイベントの確信度スコアが高い順に並び替えて K 個選択する．N は十分に大きい数が設定される ($N = 100$)．なお，K も予測対象である．

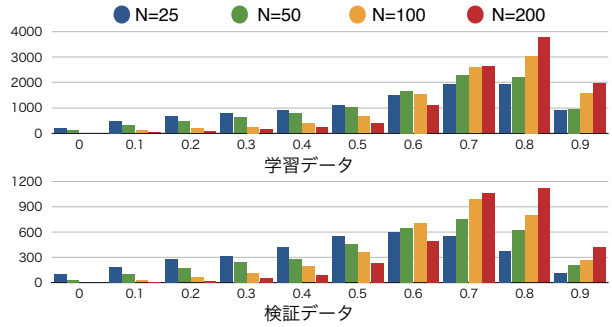


図 2 オラクルにより選択されたイベントの tIoU のヒストグラム．

3.1 イベント選択器

イベントのエンコード．イベント選択器は，入力のイベント系列 $\boldsymbol{X}$ の中から正しいイベントを選択することを目的とする．イベント選択器は Event Encoder と Event Transformer の 2 つのモジュールからなる．Event Encoder は入力の各イベント $\boldsymbol{x}_n$ を中間表現 $\boldsymbol{e}_n$ へ変換する．本研究では，Event Encoder として，映像の事前学習モデルの 1 つである Howto100M [12] によって学習済みの MIL-NCE モデル [13] を用いる．

イベントの選択．次に，イベント間の関係性を考慮するために，$\boldsymbol{e}_n$ に Positional Encoding (PE) と各イベントの情報 (開始時間，終了時間) を埋め込んだベクトルと共に加算した後，Event Transformer へ入力する．この Event Transformer は MART [14] を基にしており，前に選択したイベントの情報を記憶する役割を持つメモリベクトル $\boldsymbol{V}_t^l$ が Transformer の l 層ごとに入っている点が通常の Transformer と異なっている．Event Transformer の出力 $\boldsymbol{h}_n$ とメモリベクトル $\boldsymbol{V}_t^l$ を用いて，以下のように n 番目のイベントが選択される確率 $p(n|\boldsymbol{V}_t^l, \boldsymbol{X})$ を計算する．

$$\boldsymbol{V}_t = \max(\boldsymbol{V}_t^1, \ldots, \boldsymbol{V}_t^l, \ldots, \boldsymbol{V}_t^L) \quad (1)$$

$$p(n|\boldsymbol{V}_t^l, \boldsymbol{X}) = \frac{\exp\left((\boldsymbol{h}_n^t)^T \boldsymbol{V}_t\right)}{\sum_i \exp\left((\boldsymbol{h}_i^t)^T \boldsymbol{V}_t\right)} \quad (2)$$

ここで，max(·) はベクトルの要素ごとの最大値を取る．イベント選択器は，学習時は Gumbel softmax [15] を用いてイベントを 1 つ選択する．これにより，イベントの選択を微分可能なまま処理でき，End-to-end の学習が可能となる．推論時は $p(n|\boldsymbol{V}_t^l, \boldsymbol{X})$ の値が最大となるイベントを選択して文生成器へ送る．

3.2 文生成器

文生成器は選択されたイベントに対応する文を生成する．ここで，選択されたイベントの添字を $\hat{c}_t$ とすると，イベントベクトルは $\boldsymbol{h}_{\hat{c}_t}$ と書ける．文生成器は，まず t 番目のイベントに対応する文 $\boldsymbol{y}_t$ の単語列を訓練済みの GloVe [16] を介して単語の分散表現へ変換する．次に，これらを $\boldsymbol{h}_{\hat{c}_t}$ の次元数と同じになるように 2 層の多層パーセプトロンによって変換し，$\boldsymbol{h}_{\hat{c}_t}$ および PE と加算する．最後に，Event Transformer と同様にメモリを持つ Senence Transformer へ入力し，その出力を softmax 層を通すことで出力単語の条件付き確率分布を得る．

3.3 メモリの更新

Event Transformer と Sentence Transformer のメモリモジュール $\boldsymbol{V}_t^l, \boldsymbol{S}_t^l$ は別々に更新することも可能である．しかし，直感的には，相互的に情報を共有することが映像全体で一貫したストーリーの出力を得るのに有効だろうと考えられる．これを実現するために，我々はメモリモジュールを以下の式に則って更新する．

$$\hat{\boldsymbol{V}}_t = f_1(\boldsymbol{V}_t) \odot \sigma(g_2(g_1(\boldsymbol{S}_t))) \quad (3)$$

$$\hat{\boldsymbol{S}}_t = g_1(\boldsymbol{S}_t) \odot \sigma(f_2(f_1(\boldsymbol{V}_t))) \quad (4)$$

ここで，$f_*(\cdot), g_*(\cdot)$ は 1 層の線形層，$\odot, \sigma$ はアダマール積およびシグモイド関数を表す．得られた $\hat{\boldsymbol{V}}_t$ と $\hat{\boldsymbol{S}}_t$ をそれぞれの新たなメモリベクトルとして $t+1$ 番目の処理に活用する．

3.4 学習

2 つの損失関数の和を計算してモデル全体を学習させる．1 つ目は，イベント選択器を学習させるためのもので，以下の式で表される負の対数尤度 L_e を最小化する．

$$L_e = -\sum_{(\boldsymbol{X},\boldsymbol{C})} \log p(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{V}_1^l, \ldots, \boldsymbol{V}_T^l) \quad (5)$$

2 つ目は，文生成器を学習させるためのもので，t 番目の文 $\boldsymbol{y}_t$ に対し以下の負の対数尤度 L_s を最小化する．

$$L_s = -\sum_{(\boldsymbol{X},\boldsymbol{C},\boldsymbol{Y})} \sum_t \log p(\boldsymbol{y}_t|\boldsymbol{h}_{\hat{c}_t}, \boldsymbol{S}_{<t}^l) \quad (6)$$

4 実験

データセット．実験では，学習および評価のデータセットとして，YouCook2 [9] を利用した．この

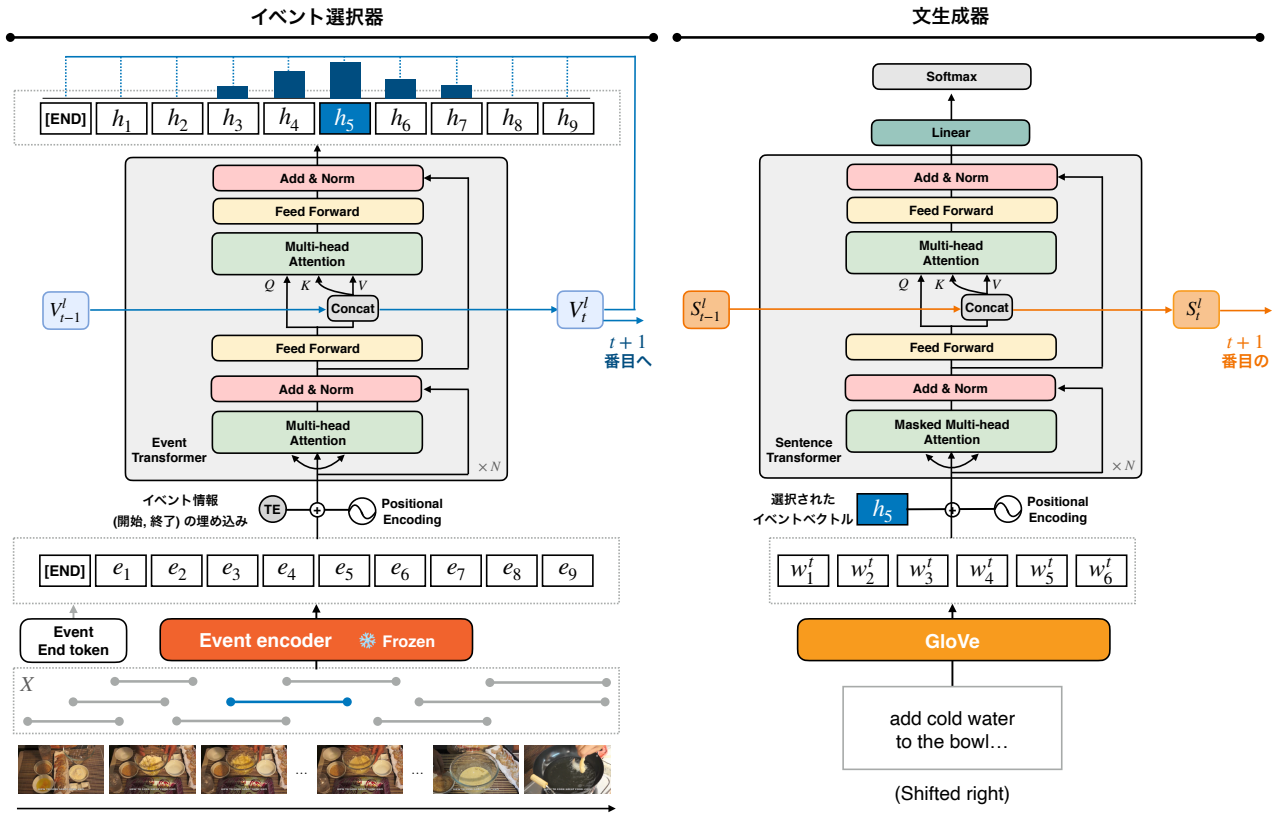


図 3　提案手法の概要図.

表 3　提案手法と比較モデルの DVC 評価結果.

	dvc_eval			SODA	
	BLEU4	METEOR	CIDEr-D	METEOR	CIDEr-D
Masked Transformer	0.46	4.09	7.62	0.36	1.32
PDVC	0.89	4.52	21.50	3.98	15.10
提案手法					
別々に更新	**1.02**	5.59	22.79	5.08	22.14
共有して更新	0.97	**5.76**	**23.78**	**5.25**	**25.30**

データセットは，YouTube 上の料理動画に対してイベントと対応する文のアノテーションがされたデータセットである (イベントの数は 3 から 16). YouCook2 ではテストセットが公開されていない. 先行研究との公平な比較を行うために，以下の結果は全て検証セットでの評価結果である.

比較モデル. 比較モデルとして，提案手法と同様に Transformer をベースとした DVC 手法である Masked Transformer [6] と PDVC [7] を採用した. また，イベントの候補数 N は 100 として設定した.

4.1　定量評価

表 3 に評価結果を示す. 提案手法はベースラインの手法よりも高い性能を，とりわけ SODA において顕著な差が見られた. また，メモリの更新方法を比較すると，共有して更新する方が別々に更新するよりも概ね高い性能を示すことも分かった.

4.2　定性評価

図 4(付録) に Ground Truth に加えオラクル，PDVC，提案手法の予測結果を示す. PDVC はストーリー性を無視したイベント予測を行っているのに対し，提案手法は動画全体を考慮して，イベント間の重なりを少なくかつ正しく予測できていることが分かる. また，出力文を比較しても，提案手法はいくつかの材料の認識の失敗はあるが，ストーリー性に沿った文章を生成できていることが分かる (混ぜる → 入れる → 浸ける → 揚げる).

5　まとめ

本研究では，映像からストーリーを生成することを目的とする. 既存の DVC モデルの出力を分析したところ，出力からうまくイベントを選べば正しいストーリーを得られることが分かった. この分析をもとに，イベントを選択するイベント選択器と文を生成する文生成器を同時学習するモデルを提案した. 実験では，提案手法が既存手法と比較して正しくストーリーを生成できていることが分かった.

謝辞

本研究は JSPS 科研費 JP21J20250 の助成を受けたものです.

参考文献

[1] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. In **Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)**, 2020.

[2] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2015.

[3] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In **Proceedings of the European Conference on Computer Vision (ECCV)**, 2020.

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In **Proceedings of the International Conference on Computer Vision (ICCV)**, 2017.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)**, 2017.

[6] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2018.

[7] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In **Proceedings of the International Conference on Computer Vision (ICCV)**, 2021.

[8] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation flamework. In **Proceedings of the European Conference on Computer Vision (ECCV)**, 2020.

[9] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In **Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)**, 2017.

[10] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, 2005.

[11] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2015.

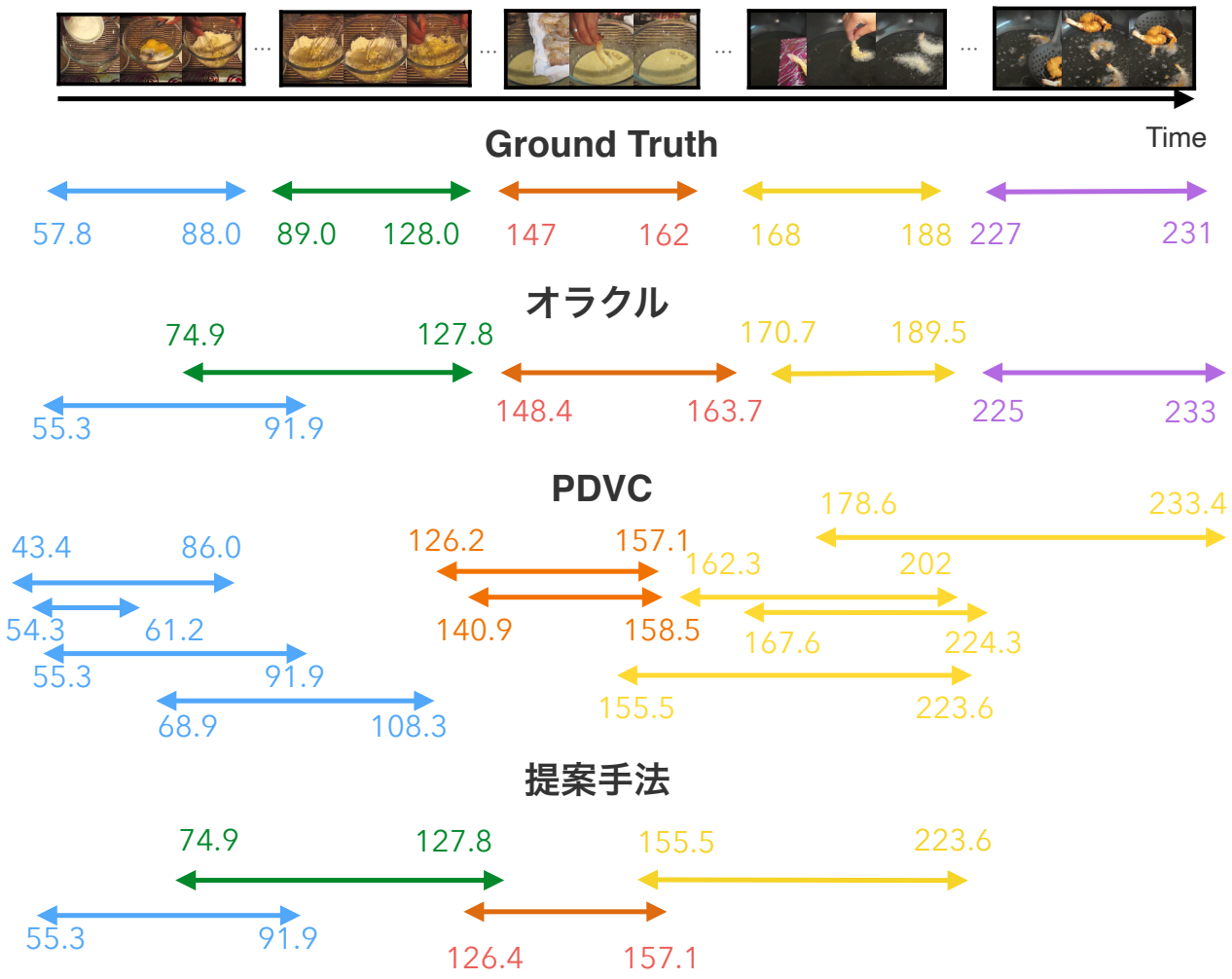
[12] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In **Proceedings of the International Conference on Computer Vision (ICCV)**, 2019.

[13] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2020.

[14] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In **Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)**, 2020.

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In **The International Conference on Learning Representations (ICLR)**, 2017.

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In **Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)**, 2014.



Ground Truth

(1) Add flour eggs baking soda salt and pepper to the bowl and stir
(2) Add cold water to the bowl and stir
(3) Cover the shrimp in the batter and breadcrumbs
(4) Place the shrimp into a pan of hot oil
(5) Remove the shrimp from the pan

オラクル

(1) Add flour and mix to a bowl and mix
(2) Mix the ingredients in the bowl
(3) Place the chicken in the batter
(4) Fry the chicken in the oil
(5) Remove the soup from the oil

PDVC

(1) Mix flour and pepper in a bowl and mix
(2) Coat the chicken in the flour mixture
(3) Fry the chicken in the oil

提案手法

(1) Mix flour pepper and beer together
(2) Add flour and corn flour to the bowl and mix Place the chicken in the batter
(3) Coat the shrimp in the flour and the batter
(4) Fry the fish in the oil

図 4 Groud Truth，オラクル選択，PDVC，そして提案手法によるイベントの予測結果およびストーリーの比較.