大規模視覚言語モデルを用いた日常生活動画からのイベント中心知識グラフ生成

後藤 颯志 [1,2]　チャクラボルティ シュデシナ [1]　森田 武史 [1,2] 太田 葵 [2]
Imrattanatrai Wiradee[2] 浅田 真生 [2] 江上 周作 [2] 鵜飼 孝典 [2,3] 濱崎 雅弘 [2]
[1] 青山学院大学　[2] 産業技術総合研究所　[3] 富士通株式会社
morita@it.aoyama.ac.jp　masahiro.hamasaki@aist.go.jp

概要

【背景と目的】本研究では，大規模視覚言語モデル（Large Vision-Language Model, LVLM）と大規模言語モデル（Large Language Model, LLM）を連携させ，日常生活動画からイベント中心知識グラフを自動生成する手法を提案する．

【手法】まず，動画を短時間クリップに分割し，LVLM を用いて箇条書き形式のキャプションを生成する．次に，LLM を用いてキャプションからトリプルを抽出し，行動情報を構造化する．抽出したトリプルに対して，VirtualHome の語彙に基づく語彙制約と埋め込み類似度により動作語・物体語を補正する．さらに，不自然行動や重複行動の削除などのルールベース処理，および LLM により時系列整合性を検証し，イベント列を洗練する．

【結果と考察】VDAct に含まれる 1,000 本の動画を対象に評価実験を実施した．その結果，提案手法はトリプル抽出精度，イベント順序整合性においてベースライン手法を上回り，その有効性が示された．

1　はじめに

動画理解は監視，HCI，医療，教育など幅広い分野で重要視されている [1, 2]．動画には膨大な視覚・時間情報が含まれるが，生のデータ形式ではコンピュータによる高度な意味処理が困難である．そのため，人，物体，および行動の相互関係を構造的に表現する枠組みが求められている．動画情報を構造化することにより，対象の意図推定や因果関係の把握が可能となり，Video QA，行動予測，動画検索などの高度なタスクへの応用が期待される．

動画の構造的表現手法として，シーングラフ，知識グラフ（Knowledge Graph, KG）が提案されている [3, 4]．これらの手法は物体間の空間的な関係表現に優れる一方で，イベント間の時間的・因果的な関係を記述する能力には限界がある．特に，複数のイベントが連続的かつ並行的に発生する日常生活動画においては，イベント単位での関係抽出は依然として困難な課題である．

一方，近年の大規模視覚言語モデル（Large Vision-Language Model, LVLM）の発展により，動画キャプション生成の精度が向上している [5, 6, 7]．その結果，自然言語による記述を基点とした知識グラフ構築は，実用可能な段階に達しつつある．

本研究では，家庭シミュレータ VirtualHome [8] から得られた日常生活動画を対象とし，LVLM と大規模言語モデル（Large Language Model, LLM）を連携させ，イベント中心知識グラフを自動生成する手法を提案する．提案手法の有効性を検証するため，日常生活動画データセット VDAct [9] に含まれる 1,000 本の動画を対象に評価実験を実施する．

2　関連研究

2.1　VirtualHome2KG

VirtualHome2KG [10] は，VirtualHome [8] のシミュレーション結果から行動と物体状態をイベント単位で整理し，知識グラフとして構造化する手法である．同手法はイベント中心の設計を採用し，各行動の内容だけでなく，時間的・因果的関係を明示的に記述できる点に特徴がある．これにより，単発の行動だけでなく，複数行動が連続・依存して発生する生活シーンをより自然に表現できる．

従来のシーングラフは物体間の関係表現には適しているが，イベント間の関係を直接的に扱うことは困難である．これに対し，VirtualHome2KG はイベントを独立したノードとして定義することで，複数の

イベントにおける連続性や依存関係を表現できる．

本研究では，このイベント中心構造を基盤とし，LVLM が生成したキャプションからイベント中心知識グラフを自動生成する手法の有効性を検証する．

2.2 VDAct

VDAct [9] は，VirtualHome [8] における家庭内行動シーンを対象としたマルチモーダルデータセットである．各動画には，人手によってアノテートされた対話データおよびイベント中心知識グラフが付与されている．この知識グラフは VirtualHome2KG [10] に基づいて構築されており，イベント属性および時間的・因果的関係を含む構造をもつ．

VDAct は動画と知識グラフが直接対応している点に特徴があり，キャプション生成やトリプル抽出の精度を定量的に評価するための基盤を提供する．本研究では VDAct を評価用データセットとして採用する．LVLM によるキャプション生成，およびトリプル抽出手法の性能をベースライン手法と比較することで，イベント中心知識グラフの自動構築における有効性と課題を検証する．

3 提案手法

3.1 提案手法の概要

本章では，動画データから日常行動の文脈を理解し，イベント中心知識グラフを自動構築する手法について述べる．提案手法の構成を図 1 に示す．

提案手法は，動画キャプション生成，トリプル抽出，トリプルの洗練，およびイベント中心知識グラフの構築の四つのプロセスで構成される．以下では，各プロセスの内容を説明する．

3.2 動画キャプション生成

本プロセスでは，LVLM 用いて動画キャプションを生成する．対象動画は長尺であるため，10 秒単位に分割し，区間境界での文脈欠落を防ぐ目的で 3 秒のオーバーラップを設定する．10 秒区間は，VirtualHome2KG における一連の行動（walk → open など）を包含することを想定した長さであり，3 秒のオーバーラップは，処理コストを考慮した設定である．キャプション形式は，後段処理との整合性を高めるため，箇条書きの短文形式とする．さらに，使用される動作語と物体語の揺れを抑制するため，VirtualHome で定義されているアクションリストおよびオブジェクトリストを事前に提示し，語彙をその範囲に限定した．これらのリストは，本研究で最終的に構築するイベント中心知識グラフにおける標準語彙としても採用しており，キャプション生成段階から一貫した語彙体系を適用することで，後段のトリプル抽出や正規化処理を容易にしている．

3.3 トリプル抽出

本プロセスでは，生成したキャプションを LLM に入力し，「(man, action, object)」形式のイベントトリプルを抽出する．なお，主語は，VirtualHome のエージェント定義に従い， man で固定している．抽出語彙には，提示したリストの範囲内に収まるよう制約を与え，言い換えや曖昧な表現を抑制する．その後，小文字化や記号除去などの正規化を行い，不一致となる語彙に対しては，LLM が出力した語の埋め込みベクトルと VirtualHome に基づくアクション語彙およびオブジェクト語彙の埋め込みベクトルとのコサイン類似度に基づき，最も近い候補語にマッピングする語彙補完を適用する．また，目的語を持たない動作には none を割り当て，機械処理に適した形式へ統一する．

3.4 トリプルの洗練

本プロセスでは，抽出されたトリプルには冗長な動作や不自然な行動が含まれるため，ルールベース処理と LLM を用いて二段階で洗練する．

1. ルールにより同一操作の連続や，テーブル等への不自然な着座動作など，日常生活の常識に反するイベントを削除する．
2. LLM によりイベント列の時系列整合性を検証し，必要な移動行動の補完や冗長なイベントを削除する．

これらの処理により，自然な生活行動に近いイベント列へ再構成できる．

3.5 イベント中心知識グラフの構築

本プロセスでは，イベント中心スキーマに基づいて，洗練されたトリプルをイベントノードとして実体化し，各ノードに agent，action，object を属性として付与する．さらに，動画内での出現順に基づき，ノード間に nextEvent 関係を付与することで，行動の流れを時系列構造として表現する．このイベント中心スキーマにより，従来のエージェントの状態変

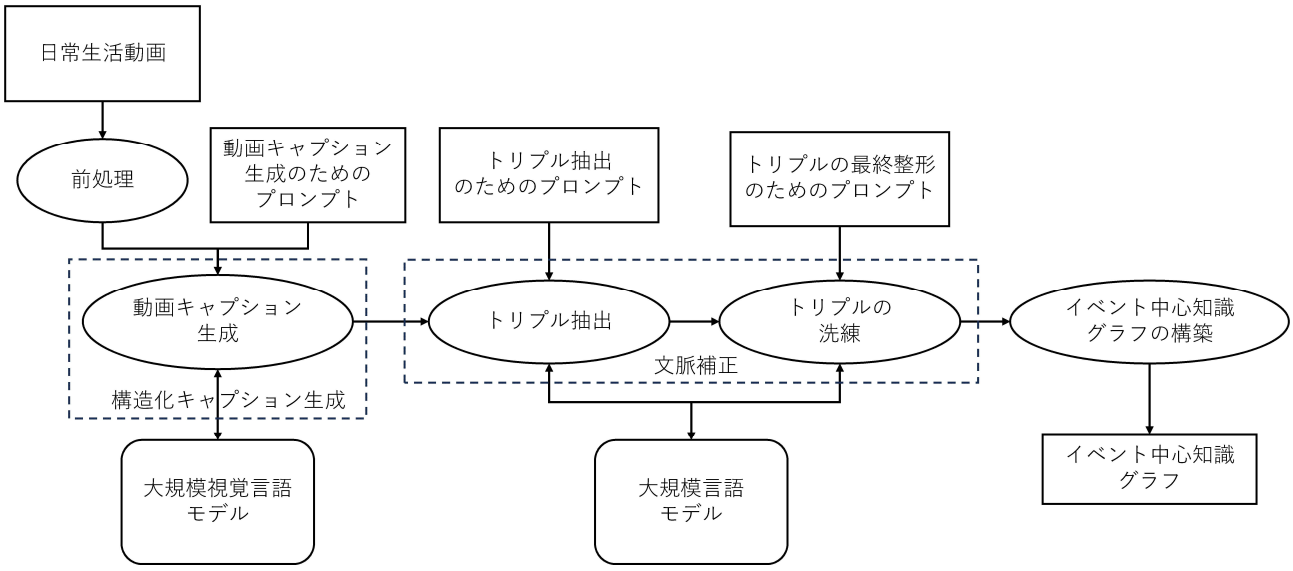

図 1　提案手法の構成

化を主軸とするエージェント中心スキーマでは表現が困難であった連続性や因果関係を含む構造的知識グラフを構築できる．

4　評価実験

4.1　実験設定

本研究では，VDAct [9] に含まれる正解知識グラフと，提案手法により構築した知識グラフを比較し，提案手法の有効性を検証する．評価対象は RDF 形式の 知識グラフである．

ベースラインには，動画キャプションと基本的な自然言語処理に基づいてイベント抽出を行う手法 [11] を用いる．この手法では，VideoLLaMA 3 [5] により生成されたキャプションに対して spaCy [12] を用いて述語項を抽出する．次に，語彙類似度に基づく補完によってトリプルを生成し，LLM による簡易的な妥当性確認を行う．本実験では，このベースライン手法と提案手法を比較することで，抽出形式の変更および洗練処理の効果を評価する．

評価指標には，適合率，再現率，F1 値を用いる．評価は，(1) 行動と対象物の一致，(2) 行動一致，(3) 対象物一致の三つの観点から行う．さらに，時系列構造の妥当性を評価するため，抽出されたイベント順序に対して編集距離を算出する．

入力動画には VDAct の日常生活動画 1,000 本を使用した．キャプション生成には VideoLLaMA 3 を，トリプル抽出および洗練には GPT-4o mini を用いた．さらに，アクション・オブジェクトの類似語補完に

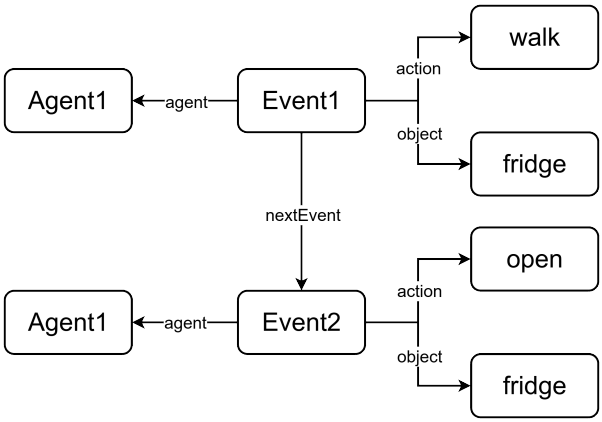

図 2　イベント中心知識グラフの例

は OpenAI の text-embedding-3-large [13] を使用した．

4.2　評価用データセット

評価には，VDAct に付与されたイベント中心知識グラフを用いる．VDAct の知識グラフは VirtualHome2KG を基盤として構築されており，各イベントに関連する主要オブジェクトおよび対象オブジェクトが整理されている．イベント中心知識グラフの例を図 2 に示す．この例は，エージェントが冷蔵庫に向かって歩き，冷蔵庫を開けるという一連の行動を表現したイベント中心知識グラフである．

4.3　実験結果と考察

本節では，VDAct を対象に，ベースライン手法と提案手法を比較した結果について述べる．表 1 にベースライン手法と提案手法の比較結果を示す．

表 1　提案手法とベースラインの比較結果

評価対象	評価指標	ベースライン手法	提案手法
行動と対象物の組	平均適合率	0.099	**0.259**
	平均再現率	0.044	**0.379**
	平均 F1 値	0.060	**0.300**
行動	平均適合率	0.518	**0.535**
	平均再現率	0.230	**0.771**
	平均 F1 値	0.313	**0.615**
対象物	平均適合率	0.245	**0.336**
	平均再現率	0.109	**0.491**
	平均 F1 値	0.148	**0.388**
イベント順序	編集距離	0.166	**0.376**

4.3.1　行動と対象物の組の評価

行動と対象物の組の評価において，提案手法は F1 値 0.300 を示し，ベースライン手法の 0.060 を上回った．また，提案手法では再現率の向上が顕著であり，ベースライン手法の 0.044 に対して 0.379 を示した．これは，短文形式のキャプションおよび語彙制約により抽出対象が明確化され，さらに LLM による補完と整合性検証の導入したことで，動画内の行動情報を網羅的に抽出できたためである．

4.3.2　行動の評価

行動の評価では，提案手法は F1 値 0.615 を示し，ベースライン手法の F1 値 0.313 を上回った．また，再現率は 0.771 であり，ベースライン手法の 0.230 より高い値を示した．これらの結果は，LLM によるトリプル抽出と語彙制約の組み合わせた提案手法が，動画内に含まれる多様な動作表現を安定して検出できていることを示す．一方，ベースライン手法は spaCy による述語抽出に依存しているため，複雑な文構造や省略を含む表現への対応が十分ではなかったと考えられる．

4.3.3　対象物の評価

対象物の評価においても，提案手法は F1 値 0.388 を示し，ベースライン手法の 0.148 を上回った．また，再現率は 0.491 であり，ベースライン手法の 0.109 と比較して高い値を示した．この結果は，語彙リストを事前に提示したことによって物体名の表記揺れが抑制されたことを示している．さらに語彙補完およびルールベース処理が曖昧な対象語を正規化したことで，行動対象となる物体の認識精度が向上した．

4.3.4　イベント順序の比較

イベント順序に対する編集距離の評価において，提案手法は 0.376 を示し，ベースライン手法の 0.166 を上回った．この結果は，提案手法が行動の前提条件となる移動行動などを LLM により補完し，冗長な行動を削除したことで，実際の生活行動に近いイベント列を再構築できたためである．

4.3.5　考察

以上の結果より，提案手法は，トリプル抽出，行動語抽出，対象物抽出，および時系列構造の各観点において，ベースライン手法を上回る性能を示した．特に，再現率とイベント順序の整合性が改善されており，その要因として以下の点が挙げられる．

1. キャプションを短文形式とし，語彙制約を導入することで，抽出対象が明確な構造化テキストを生成できた点．
2. トリプル抽出を LLM ベース手法へ移行したことで，複雑な動作や文脈に依存する表現への対応が可能となった点．
3. ルールベース処理と LLM による洗練により，冗長イベントや非現実的な行動を除去し，時系列整合性を維持できた点．

これらの結果は，提案手法が動画理解における行動抽出およびイベント構造化の精度向上に有効であることを示している．

5　おわりに

本研究では，LVLM と LLM を連携させ，日常生活動画からイベント中心知識グラフを自動構築する手法を提案した．短文形式のキャプション生成，LLM によるトリプル抽出，およびルールベース処理と LLM を併用した洗練により，動画中の行動を一貫したイベント列として再構成できることを示した．

VDAct に含まれる動画 1,000 本を対象とした評価では，提案手法はベースライン手法と比較して，トリプル抽出精度およびイベント順序の時系列整合性が向上した．これらの結果から，構造化キャプション生成と文脈補正が，イベント抽出精度の向上に寄与することが確認された．

今後の課題として，因果関係や状態遷移を含む高次の知識抽出，複数エージェント推論，ならびに実世界動画への適用に向けた頑健性の検証が挙げられる．

謝辞

本研究成果の一部は，国立研究開発法人新エネルギー・産業 技術総合開発機構 (NEDO) の委託業務 (JPNP20006，JPNP25006) の結果得られたものです．本研究は JSPS 科研費 23K11221，25K03232 の助成を受けたものです．

参考文献

[1] Yuan-Heng Lin, Yi-Hsing Chien, Min-Jie Hsu, Wei-Yen Wang, Cheng-Kai Lu, and Chen-Chien Hsu. A Data Monitoring System with Human Action Recognition for Long-Term Care Institutions. In **2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)**, pp. 51–52, 2023.

[2] Lin Guo, Zongxing Lu, and Ligang Yao. Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review. **IEEE Transactions on Human-Machine Systems**, Vol. 51, No. 4, pp. 300–309, 2021.

[3] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 2856–2865, June 2021.

[4] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition and applications. **arXiv preprint arXiv:2002.00388**, 2020.

[5] Boqiang Zhang, et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. **arXiv, abs/2501.13106**, 2025.

[6] An Yang, Anfeng Li, Baosong Yang, and Beichen Zhang+. Qwen3 technical report. Technical report, arXiv preprint arXiv:2505.09388, 2025.

[7] Jinguo Zhu, Weiyun Wang, Zhe Chen, and Zhaoyang Liu+. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. Technical report, arXiv preprint arXiv:2504.10479, 2025.

[8] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating Household Activities via Programs. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2018.

[9] Wiradee Imrattanatrai, Masaki Asada, Kimihiro Hasegawa, Zhi-Qi Cheng, Ken Fukuda, and Teruko Mitamura. A video-grounded dialogue dataset and metric for event-driven activities. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 39, No. 23, pp. 24203–24211, 2025.

[10] Shusaku Egami, Takanori Ugai, Mikiko Oono, Koji Kitamura, and Ken Fukuda. Synthesizing event-centric knowledge graphs of daily activities using virtual space. **IEEE Access**, Vol. 11, pp. 23857–23873, 2023.

[11] 後藤颯志, チャクラボルティシュデシナ, 森田武史, 太田葵, Imrattanatrai Wiradee, 浅田真生, 江上周作, 濱崎雅弘, 鵜飼孝典. 大規模視覚言語モデルを用いた日常生活動画からのイベント中心知識グラフ生成の検討. 人工知能学会第二種研究会資料, Vol. 2025, No. SWO-066, p. 05, 2025.

[12] Matthew Honnibal and Ines Montani. spacy: Industrial-strength natural language processing in Python. https://spacy.io/, 2020.

[13] OpenAI. text-embedding-3-large: Openai embedding model. https://platform.openai.com/docs/guides/embeddings, 2024.

Frame 004　　Frame 007

図 3　動画フレームの例

```
The man walks to the television.
The man turns on the television.
```

Listing 1　図 3 に示す動画フレーム列から生成された動画キャプションの例

A　動画キャプションおよびイベント中心知識グラフ生成例

本付録では，図 3 に示した動画フレーム列に対して，提案手法が生成する動画キャプションの例およびイベント中心知識グラフの例を示す．これにより，提案手法が動画内容をどのように構造化表現へ変換するかを具体的に示す．

A.1　動画キャプション生成例

図 3 に示す動画フレーム列（Frame 004 および Frame 007）に基づき，LVLM によって生成された動画キャプションの例を Listing 1 に示す．

これらの文は，エージェント（agent），行動（action），対象物（object）といったイベント要素を明確に含んでおり，後続の述語項抽出処理へそのまま接続できる．

本研究では，動画キャプション中に含まれる動詞を VirtualHome のアクション語彙に対応付けるため，語彙補完処理を導入している．VirtualHome のアクションリストには「turn on」が含まれていない．そのため，出力された動詞の埋め込みベクトルと，VirtualHome に基づくアクション語彙集合に含まれる各語の埋め込みベクトルとのコサイン類似度を算出し，最も類似度の高い語へマッピングする．本例では，「turn on」に最も近い語として「switch on」が選択され，アクションラベルとして補完された．

A.2　イベント中心知識グラフ生成例

提案手法におけるトリプル抽出および洗練処理により，動画内の二つのイベントは，Listing 2 に示すように RDF 形式で表現される．

```
<vh2kg_url/instance/00001_event_01> <vh2kg_url/ontology/agent>
      <vh2kg_url/instance/man> .
<vh2kg_url/instance/00001_event_01> <vh2kg_url/ontology/action>
      <vh2kg_url/ontology/walk> .
<vh2kg_url/instance/00001_event_01> <vh2kg_url/ontology/object>
      <vh2kg_url/ontology/tv> .

<vh2kg_url/instance/00001_event_02> <vh2kg_url/ontology/agent>
      <vh2kg_url/instance/man> .
<vh2kg_url/instance/00001_event_02> <vh2kg_url/ontology/action>
      <vh2kg_url/ontology/switchon> .
<vh2kg_url/instance/00001_event_02> <vh2kg_url/ontology/object>
      <vh2kg_url/ontology/tv> .

<vh2kg_url/instance/00001_event_01> <vh2kg_url/ontology/NextEvent>
    <vh2kg_url/instance/00001_event_02> .
```

Listing 2　イベント中心知識グラフの例

```
ROLE: You are an expert video content analyzer specialized in
      structured knowledge extraction.

TASK: Convert video content into structured bullet-point
      descriptions suitable for knowledge graph construction.

OUTPUT REQUIREMENTS:
Format: [Subject] + [Action Verb] + [Object/Complement]
Structure: Subject-Predicate-Object triples
Language: Simple, clear, grammatically correct English
Formatting: Output one sentence per line(separated by line breaks)

AVAILABLE VOCABULARY:
Actions: [ アクションリスト ]
Objects: [ オブジェクトリスト ]
...
```

Listing 3　動画キャプション生成に用いたプロンプトの一部

本例では，「walk」と補完後の「switch on」という二つの行動が抽出され，それぞれに対応するエージェント（man）および対象物（tv）が明確に紐づけられている．さらに，時間的順序を表す `nextEvent` 関係により，動画中の行動列が知識グラフ上に再構成されている．

A.3　作成した主要なプロンプト

動画キャプション生成に用いたプロンプトの一部を Listing 3 に，トリプルの洗練に用いたプロンプトの一部を Listing 4 示す．

```
You are a temporal event consistency checker for human activities
      in a house.
Given the following event triples extracted from a video:
[ triple ]
Each triple is in the form:
(man, action, object)

Please:
1. Remove redundant or obviously repetitive actions (especially
      repeated "walk" to the same destination when nothing
      meaningful happens in between).
2. Preserve room-level movements such as moving from one room to
      another (e.g., livingroom -> kitchen -> bedroom).
3. Keep final destinations where the man performs an action (e.g.,
       sit, grab, open, close, switch on/off).
...
```

Listing 4　トリプルの洗練に用いたプロンプトの一部