# 談話構造に基づく Video Question-Answering ベンチマークデータセットの整備

吉岡智輝 [1]　平尾努 [1]
[1] 金沢大学
{y2252100242@stu,hirao@se}.kanazawa-u.ac.jp,

## 概要

動画理解において、複数のイベントが関わり合うことで形成される動画全体のストーリー構造を理解する能力は必要不可欠である。しかし、視覚言語モデル (Vision and Language Model:VLM) の動画理解能力を測るために提案された Video Question-Answering (VQA) ベンチマークは、単一のイベント区間に対する単純な質問が多く、VLM のストーリー理解能力を十分に評価できていない。本研究では、動画中のイベント間に成り立つ談話関係を手がかりとして、ストーリー構造中の因果関係などの論理構造とその根拠を問う VQA ベンチマークデータセットを構築した。我々の VQA ベンチマークを用いて、最新の VLM である Gemini 2.5 Pro、InternVL3.5-Instruct(8B および 14B)、Qwen3-VL-Instruct(8B および 32B) を評価した結果、動画ストーリー理解にまだ改善の余地が残されていることが明らかとなった。

## 1　はじめに

近年、視覚言語モデル (Vision and Language Model: VLM) の急速な発展により、計算機による動画理解能力は著しく進歩している [1, 2, 3]。VLM の能力を客観的に測るための Video Question-Answering (VQA) ベンチマークデータセットが整備されている [4, 5]。しかし、これらのベンチマークは、数秒程度の短いクリップに対し「画面に何が映っているか (物体の同定)」や「どのような動作が行われているか (動作認識)」といった、局所的かつ表層的な事実を問う単純な質問で構成される [6, 7]。一方、人が観ることを前提とした動画コンテンツには、「何か」を伝えるためのストーリーがある。つまり、動画中のイベント (動画中の区間) は個々に独立しているのではなく、それらが関わりを持つことでストーリーを形成する。こうした動画のストーリーを VLM が捉えることができているかを客観的に評価することこそ、今後の VLM の発展の方向性を決めるうえで重要である。しかし、先に述べた既存の VQA ベンチマークは、ストーリー理解能力を測るという観点では適しているとはいえない。

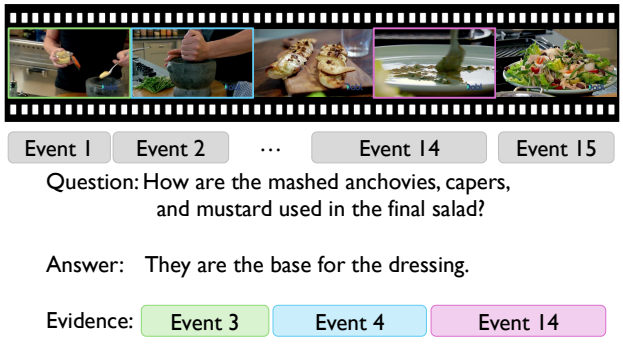


**図 1**　提案する VQA タスクの例（VideoID: 735）

本研究では、動画ストーリー理解のための枠組みである動画談話構造 [8] に着目した VQA ベンチマークデータセットを整備する。動画談話構造では、動画内の各イベントを談話の最小構成素とし、それらの間に成立する因果や条件といった論理的な関係を談話関係と呼ぶ。そして、隣接するイベント (区間) を談話関係を用いて再帰的に結びつけることで動画全体の木 (談話構造木) を構成する。これにより、時間的に離れていてもイベントの間にどのような談話関係が成り立つかがわかる。本研究では、この談話構造木に基づき、複数のイベントにまたがる関係を問う VQA ベンチマークデータセットを整備する。なお、VLM が談話構造を正しく捉えているならば、質問に対する回答だけでなく、その導出元となったイベント (談話の最小構成素) も正確に特定できるはずである。そこで本研究では、テキストによる質問と回答だけでなく、回答の根拠となったイベントも同時に提示することで、VLM が動画のストーリー構造をどの程度理解できているのか、その推論プロセスを含めて評価する。我々の VQA タス

クを図 1 に示す。VLM は動画とイベント区間のアノテーション、質問を受け取り、回答を生成するとともにその根拠となったイベントを出力する。

我々の VQA ベンチマークを用いて Gemini 2.5 Pro[9]、InternVL3.5-Instruct(8B および 14B)[10]、Qwen3-VL-Instruct(8B および 32B)[11] を評価したところ、Gemini の回答の正答率は 0.5 程度、根拠抽出精度は 0.7 程度であり、InternVL3.5、Qwen3-VL のそれらは 0.25 から 0.35、0.4 から 0.55 程度と Gemini よりは大幅に低いスコアとなった。これらより、最先端モデルであっても動画のストーリー理解にはまだ課題が残っていることが明らかとなった。

## 2 関連研究

### 2.1 VQA ベンチマーク

初期のベンチマークとしては MovieQA [12] が提案された。これは映像を題材に登場人物の意図や行動理由を問うストーリー理解を評価するものである。しかし、字幕情報への依存度が高く、映像情報の利用が限定的であることが指摘された [13]。その後、視覚的な理解を重視したベンチマークが相次いで登場した。例えば、ハウツー動画の作業工程を問う How2QA [14] や、日常生活における動作の特定、発生順序、頻度といったイベントの推移を問う ActivityNet-QA [4] が挙げられる。これらの登場により、動画における動作認識能力の評価が本格化した。

近年では、単なる動作認識にとどまらず、動画内のイベント間の因果関係や時空間的な推論を要するベンチマークも提案されている。例えば、NExT-QA [6] や STAR [7]、AGQA [15] は動画内の「なぜ (Why)」や「どのように (How)」といった論理的な関係性を問う質問を含んでいる。さらに、長時間の動画を対象としたベンチマークも提案されており、例えば Video-MME [5] では、数時間に及ぶ動画に対し、映像全体の長期的な文脈理解を問うような質問が含まれている。

### 2.2 回答根拠の特定

VQA において、モデルが正答を導出したとしても、それが動画内の正しい根拠に基づいているかは自明ではない。そのため、近年では回答の生成と同時に、その根拠となる映像内の時間区間を提示するタスクが重要視されており、これは一般に Video Temporal Grounding や、Grounded VideoQA と呼ばれる。この分野は初期の研究では、自然言語クエリ (宣言文) に対応する動画区間を特定する Video Moment Retrieval タスクとして定式化され、Charades-STA [16] や ActivityNet-Captions [17] といったベンチマークが提案された。これらは主に、特定の動作やイベントの検索を目的としていたが、より複雑な質問応答への対応が求められるようになった。これを受け、Lei ら [18] は、クエリに関連するハイライトシーンの検出と区間特定を統合したベンチマークデータセットである QVHighlights を構築した。さらに最近では、VQA タスクと根拠特定タスクが統合されたベンチマークデータセットが整備されている。例えば、前述の NExT-QA を拡張した NExT-GQA [19] では、質問に対する回答だけでなく、その根拠となる時間区間の予測が求められ、モデルの説明可能性を評価する上で重要な指標となっている。

## 3 提案する VQA ベンチマーク

### 3.1 VQA タスク

VLM は、動画 $V$、質問 $Q$ に加え、動画全体を構成する $N$ 個のイベント区間の候補集合 $E = \{e_1, e_2, \ldots, e_N\}$ (各 $e_i$ は開始時刻と終了時刻を持つ) を入力として受け取る。そして、質問 $Q$ に対する自然言語による回答 $A$ を生成すると同時に、与えられた候補の中から、その回答を導き出すために不可欠な根拠イベントの集合 $\mathcal{E} \subseteq E$ を出力する。

この VQA タスクの特徴は、あるイベント区間が表現する内容とイベント区間の間に成り立つ論理的な関係を理解できなければ正答できない点にある。具体例として、図 1 に示すサラダの調理動画 (VideoID:735) を取り上げる。質問は「How are the mashed anchovies, capers, and mustard used in the final salad? (マッシュされたアンチョビ、ケッパー、マスタードは、最終的なサラダにおいてどのように使われているか？)」である。この質問は、最後のイベントにサラダが写っていることを認識するだけでは回答することができない。なぜなら、これらの食材は調理過程ですり潰され、最後のイベント区間ではドレッシングと化しているからである。この例では、VLM には (1)Event3 にて、すり鉢に食材が投入され、Event4 で、それらがすり潰されてペースト状に変化したことをとらえ、(2) それが、遠く離れた Event14

**表 1** VQA ベンチマークの比較。Multi-Spans は回答に不連続な複数区間の参照が必要か否かを、Time Labels は回答の根拠となるシーンの時間情報の有無を示す。

| Dataset | Avg. Duration (s) | Open-Ended | Multi-Spans | Time Labels |
|---|---|---|---|---|
| MovieQA | 211.4 | ✘ | ✘ | ✔ |
| NExT-QA | 44 | ✔ | ✘ | ✘ |
| Video-MME | 1017.9 | ✘ | ✔ | ✔ |
| NExT-GQA | 39.5 | ✘ | ✘ | ✔ |
| **Ours** | 105.4 | ✔ | ✔ | ✔ |

のサラダのドレッシングとして使われていることを捉えることが要求される。

### 3.2 ベンチマークデータセット

ベンチマークデータセットを整備するためには、動画に対する談話構造のアノテーションが必要となる。そこで、YouTube の数分程度の動画に対して談話構造のアノテーションを与えたデータセットである Video Discourse TreeBank (VDTB) [8] に収録されている 1,050 件の動画を対象として、マルチモーダル大規模言語モデルである Gemini 2.5 Pro を用いることで自動的に質問、回答、根拠を生成する。

VDTB では、動画内のイベントを談話の最小単位と定義し、それらイベント間の意味的な結びつきを Rhetorical Structure Theory (RST) [20] に基づく木構造として表現している。このデータセットでは、テキストの RST と同様に、イベントとそのスパンの役割を核と衛星とし、核と衛星が対となって出現する際には、衛星から核への修飾関係を因果や補足などの単核関係、核が対となって出現する際にはリストや対比などの多核関係でイベント (スパン) の修辞関係を表す。修辞関係は 9 種あり、イベント (スパン) を核と衛星に分類しつつ、それらの間の修辞関係を決定する手続きを再帰的に行うことで、動画のストーリーを木構造として表現する。

動画と修辞構造のアノテーションをともに Gemini に与えることで、質問、回答、根拠のタプルを得る。具体的には、動画とイベント区間の開始・終了タイムスタンプ、イベント間の談話関係を Gemini に与え、回答の根拠となるイベント集合が、談話構造木上で連結した部分木を形成することや、イベント間の因果・条件・順序関係といった論理的なストーリー構造に基づく推論を必須とすることをプロンプトで指示し、質問、回答、根拠のタプルを生成する。(プロンプトの詳細は付録 A を参照) 1 動画あたり 3 つの質問、回答、根拠のタプルを生成させ、合計で 3,150 件 $(1,050 \times 3)$ のタプルを得る。

既存の VQA ベンチマークとの比較を表 1 にまとめる。動画長がやや短いものの、動画中の不連続な複数区間を対象とした自由回答形式の QA であり、回答根拠となる区間を答えるという点で既存の VQA ベンチマークデータセットとは異なる。

## 4 談話構造の有効性

動画中の不連続な複数区間を参照しなければならない質問、回答、根拠区間のタプルを得ること自体は談話構造を用いずとも可能である。そこで、談話構造を用いることで質の高いデータセットを構築できるかどうかを検証する。

### 4.1 生成される回答と根拠

前節のデータセット作成時に談話構造を与えずに生成した質問、回答、根拠のタプル (これを $D_{\text{base}}$ と呼ぶ) と談話構造を与えて生成した質問、回答、根拠のタプル (これを $D_{\text{tree}}$ と呼ぶ) の違いを調べた。

各動画から得た $D_{\text{tree}}$ の 3 つの質問と $D_{\text{base}}$ の 3 つの質問の間の類似度を BERTScore [21] を用いて計算した後、割当問題を解くことで類似度が最大となる 1 対 1 のアラインメントを決定する。そして、動画全体で平均した値で $D_{\text{tree}}$ と $D_{\text{base}}$ の類似性を評価する。根拠についても同様にアラインメントを決定し、動画全体で平均した値で類似性を評価するが、テキスト間の類似性ではなく集合間の類似性なので Jaccard 係数を用いる。

その結果、質問間の類似性に関しては BERTScore が 0.911 であった。高い類似性を示しているが、質問文は同じ動画に対して生成されたものであることからある程度似ていること、つまり、共通の単語あるいは似た単語が利用されたことが原因と考える。一方、根拠間の類似性については Jaccard 係数が 0.524 であり、$D_{\text{tree}}$ と $D_{\text{base}}$ の根拠となるイベントは半数程度が一致するレベルであり高いとは言えない。

これらの結果は談話構造を与えてデータセットを構築すると、それがない場合とは異なる質問、回答、根拠が生成されることを示している。

### 4.2 データセットの整合性

動画に対する質問、回答、根拠タプルの質 (整合性) を評価するために、Liu ら [22] が提唱する Video-and-Language Inference(VIOLIN) の枠組みを採

用した。VIOLIN は、動画クリップ (Premise) が、動画内容に関する自然言語テキスト (Hypothesis) を含意 (Entailment) するか否かを判定するタスクである。$D_{\text{tree}}$ からランダムに抽出した 100 件に対し、以下の手順で評価を行った。(1) 大規模言語モデル (LLM) を用いて質問と回答のペアを単一の平叙文に変換し、これを仮説とする。(2) 根拠として選択されたタイムスタンプに基づいて動画を切り出し、これを前提とする。(3) 最後に、この前提となる動画クリップと仮説テキストを Gemini 2.5 Pro に入力し、動画の内容が仮説を支持 (含意) しているか否かを判定させる。$D_{\text{base}}$ も同様の手順で評価を行った。

実験の結果、「含意」と判定された割合は $D_{\text{base}}$ が 83 %に対し、$D_{\text{tree}}$ は 92 %と高い値を示した。動画中の不連続な複数区間を参照しなければならない質問、回答、根拠のタプルは談話構造を用いずとも得ることができる。しかし、これらの実験結果より、談話構造を利用することで、それを用いない場合とは異なる質問、回答、根拠のタプルが得られ、なおかつ、それらの信頼性が高いことが示唆された。

# 5 実験

我々の VQA ベンチマークデータセットを用い、最先端の VLM のストーリー理解力を評価した。

## 5.1 実験設定

プロプライエタリモデルとして Gemini 2.5 Pro を、オープンウェイトモデルとして InternVL3.5-Instruct(8B および 14B)、Qwen3-VL-Instruct(8B および 32B) を対象とし、動画、質問、および候補となる全イベント区間 (ID 付き) を入力し、回答テキスト (自由形式) と根拠となるイベント ID を出力するよう指示する。そして、Yao ら [23] に倣い、質疑応答ペアの妥当性を双方向の含意関係（Textual Entailment）に基づき評価する。具体的には、まず LLM を用いて質問と回答のペアを平叙文 $(S_{\text{GT}}, S_{\text{PR}})$ へと変換する。次に予測文 $S_{\text{PR}}$ と正解文 $S_{\text{GT}}$ の間の含意関係を判定し、以下のカテゴリに分類する。

- **Superior (Sup.):** $S_{\text{PR}} \rightarrow S_{\text{GT}}$ かつ $S_{\text{GT}} \not\rightarrow S_{\text{PR}}$ が成立する場合。予測回答が正解の情報を包含し、かつ情報量において正解を上回る状態を指す。
- **Equivalent (Equ.):** $S_{\text{PR}} \leftrightarrow S_{\text{GT}}$ が成立する場合 。予測と正解が意味的に等価である状態を指す。
- **Inferior/Invalid:** $S_{\text{PR}} \not\rightarrow S_{\text{GT}}$ となる場合。正解の情報が不足している状態や、内容が誤っている状態を指す。

**表 2** $D_{\text{tree}}$ における各モデルの評価結果

| | QA Acc. | | Event Grounding | | | |
|---|---|---|---|---|---|---|
| | **Sup.** | **Equ.** | **Prec.** | **Rec.** | **F1** | **EM ( /3150)** |
| **Gemini 2.5 Pro** | .170 | .395 | .769 | .626 | .690 | 589 |
| **InternVL3.5-8B** | .077 | .172 | .398 | .432 | .414 | 58 |
| **InternVL3.5-14B** | .073 | .184 | .442 | .385 | .412 | 57 |
| **Qwen3-VL-8B** | .067 | .241 | .487 | .596 | .536 | 181 |
| **Qwen3-VL-32B** | .132 | .247 | .542 | .565 | .553 | 266 |

平叙文への変換および含意関係の判定には LLM を用いる。また、根拠選択（Event Grounding）の評価には、正解のイベント集合 $E_{\text{GT}}$ と予測イベント集合 $E_{\text{PR}}$ の F1 スコア、および全イベントが一致した割合である Exact Match (EM) を用いる。

## 5.2 結果と考察

Gemini の Sup. と Equ. の合計は約 57%であり、InternVL3.5 と Qwen3-VL のそれはそれぞれ約 25%、35%に留まった。根拠特定の F 値は、Gemini は約 70%、InternVL3.5 が約 41%、Qwen3-VL が約 55%であった。オープンウェイトモデルを比較すると、Qwen3-VL が明らかに良いが、高いスコアとは言えず、Gemini との差も非常に大きい。つまり、これらのモデルが動画のストーリーを理解できているとは言い難い。

一方、Gemini は根拠特定において比較的高いスコア (69%) を達成しているものの、回答の正解率が 57%程度であることを考えると、まだまだ改善の余地が残っていることを示唆する。最先端のプロプライエタリモデルであっても我々の VQA では決して高いスコアを達成できていないことは、現状の VLM の事前学習では、動画ストーリーを理解することが困難であることが示唆される。

# 6 おわりに

本研究では、動画のストーリー理解を評価するための VQA ベンチマークデータセットを構築した。既存のデータセットである VDTB の談話構造情報を活用し、高度な推論を要する QA および回答根拠 (Evidence) の整備をした。実験の結果、最新の VLM であっても、回答精度および根拠特定において依然として課題があることが明らかになった。

## 謝辞

本研究の一部は JSPS 科研費 25K03191 の助成を受けたものです。本研究では、東京大学、情報基盤センターのスーパーコンピュータ Miyabi および産総研及び AIST Solutions が提供する ABCI 3.0 を利用した。

## 参考文献

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[2] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. **IEEE Transactions on Circuits and Systems for Video Technology**, pp. 1–1, 2025.

[4] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, No. 01, pp. 9127–9134, Jul. 2019.

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In **CVPR**, 2025.

[6] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 9777–9786, June 2021.

[7] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In **Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)**, 2021.

[8] Tsutomu Hirao, Naoki Kobayashi, Hidetaka Kamigaito, Manabu Okumura, and Akisato Kimura. Video discourse parsing and its application to multimodal summarization: A dataset and baseline approaches. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 9943–9958, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Google Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[10] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

[11] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-vl technical report. **arXiv preprint arXiv:2511.21631**, 2025.

[12] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2016.

[13] Bhavan Jasani, Rohit Girdhar, and Deva Ramanan. Are we asking the right questions in movieqa? In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops**, Oct 2019.

[14] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In **Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)**. NeurIPS, 2018.

[15] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, 2021.

[16] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In **2017 IEEE International Conference on Computer Vision (ICCV)**, pp. 5277–5285, 2017.

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

[18] Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, **Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual**, pp. 11846–11858, 2021.

[19] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 13204–13214, June 2024.

[20] WILLIAM C. MANN and SANDRA A. THOMPSON. Rhetorical structure theory: Toward a functional theory of text organization. **Text - Interdisciplinary Journal for the Study of Discourse**, Vol. 8, No. 3, pp. 243–281, 1988.

[21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[22] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2020.

[23] Peiran Yao and Denilson Barbosa. Accurate and nuanced open-QA evaluation through textual entailment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 2575–2587, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

# A　VQA の自動生成

動画に対する質問、回答、根拠のタプルを得るために用いたプロンプトは以下の通りである。

```
# Video QA Generation Task: Creating Story-Understanding Question-Answer Pairs

## Task Overview
Generate up to 3 high-quality question-answer pairs that require understanding the video's story structure through its dependency relationships. These questions must demonstrate complex reasoning across multiple connected events that form the video's discourse structure.

## Input Data
- **Video**: A video containing a story with multiple events
- **Timestamp**: [start, end] pairs are provided for each event segments in the video.the video is constructed of {NUM} segments.
{TIME}
- **Dependency Structure**: Event relationships in format: `[modifier_start, modifier_end] [head_start, head_end] relation_type`
{DEP}

## Core Requirements

### Question Constraints

1. **Story-Level Reasoning**: Questions must require understanding how events connect to form the video's narrative
2. **Connected Events Only**: All selected events must form a connected subtree in the dependency structure
3. **Multi-Event Synthesis**: Requires information from >=2 events that cannot be answered from any single event
4. **Advanced Reasoning Types**:
   - **Causal**: "How did X cause Y?" or "Why did X lead to Y?"
   - **Temporal**: "Why did X happen after Y?" or "What enabled X following Y?"
   - **Sequential**: "How did the sequence from X to Z unfold?"
   - **Enabling**: "What made X possible?" or "How did Y enable Z?"

### Quality Standards

- **Conciseness**: Questions <=22 words, answers <=10 words
- **Story Focus**: Questions should reveal understanding of the video's narrative structure
- **Dependency-Driven**: Leverage the dependency relationships to create meaningful reasoning chains

## Dependency Structure Guide

### Format
```
[modifier_start, modifier_end] [head_start, head_end] relation_type
```

### Key Relations for Video Stories
- **Cause**: Modifier event causes head event (unexpected causality)
- **Result**: Head event results from modifier (procedural outcomes)
- **Preparation**: Modifier prepares for head event (setup relationships)
- **Background**: Modifier provides context for head event
- **Supplement**: Modifier emphasizes head event (common in edited videos)
- **ROOT**: Central event of the story

### Usage for Question Generation
- **Trace dependency chains** to understand story flow
- **Connect temporally distant but logically related events**
- **Ensure all selected events form a connected subtree**
- **Focus on ROOT events as story anchors**

## Questions to Avoid

**Single-Event Questions:**
- :x: "What is the person doing?" (descriptive, single event)
- :x: "Where does X go?" (simple observation)

**Surface-Level Questions:**
- :x: "Who helps X?" (direct observation, no reasoning)
- :x: "What color is X?" (visual description only)

## Good Question Examples

:white_check_mark: **Sequential Reasoning**: "How did the preparation steps enable the final outcome?"
- Connects: [Preparation events] → [Intermediate steps] → [Final result]

:white_check_mark: **Causal Chain**: "Why did the initial setup lead to the later complication?"
- Connects: [Setup] → [Intermediate cause] → [Complication]

:white_check_mark: **Enabling Relationship**: "What made the resolution possible after the earlier failure?"
- Connects: [Failure] → [Enabling event] → [Resolution]

## Output Format

For each question-answer pair, use exactly this format:
question: <your question>
answer: <your answer>
evidence: [<comma-separated list of timestamp segment numbers used as evidence>]

### Timestamp Rules
- **Connected Events**: All timestamps must correspond to events forming a connected subtree
- **Exact Values**: Use the original event numbers provided in the Input Data
- **Complete Evidence**: Include all events necessary to justify the answer
- **Dependency Chain**: Timestamps should reflect the reasoning path through dependencies

## Generation Process

1. **Analyze dependency structure** to identify story flow and key relationships
2. **Find connected event groups** that form meaningful reasoning chains
3. **Create questions** that require understanding these connections
4. **Verify connectivity** - ensure all selected events form a subtree
5. **Check quality** against prohibited questions and requirements

## Quality Checklist

[] Requires understanding of video's story structure?
[] All selected events form a connected subtree?
[] Requires >=2 events that cannot be answered individually?
[] Uses causal/temporal/sequential/enabling reasoning?
[] <=22 words (question) and <=10 words (answer)?
[] Different from any prohibited questions?
[] Timestamps are exact copies from input?

Generate question-answer pairs that demonstrate deep story understanding through dependency-aware reasoning.
```