動画キャプション生成における マルチモーダル LLM のハルシネーション分析

仲田勝太　近藤雅芳
LINE ヤフー株式会社
{shota.nakada,masayoshi.kondo}@lycorp.co.jp

概要

本研究は、動画キャプション生成タスクにおけるマルチモーダル LLM のハルシネーションの分析を行う。また、ハルシネーション分析に向けて、モデル生成文に対する誤り区間のスパンとそのスパンの修正を付与した新しいデータセットを構築した。分析は言語情報と視覚情報の２つの観点から行い、誤りパターンの分析や動画の明るさや動きといった視覚的要因との関連性に注目した分析を行う。分析結果から、モデルの生成文は入力動画の視覚情報よりもマルチモーダル LLM の Decoder の影響を強く受けた文表現となることや、動画の明るさが大きいほどハルシネーションが生じやすい傾向にあることなどが分かった。

1　はじめに

YouTube、TikTok をはじめとする動画共有サービスや Netflix などの動画配信サービスは、日常生活に欠かせない存在となっており、日々膨大な動画コンテンツを我々に届けている。動画キャプション生成 (Video Captioning) [1, 2, 3, 4] は、そのような動画コンテンツへのタグ付けや説明文を付与できる技術であり、膨大な動画コンテンツを好みに応じて迅速に我々に届けることを可能にする実用上重要な技術のひとつである。近年は、動画キャプション生成タスクにマルチモーダル LLM [3, 4, 5, 6, 7, 8] が用いられつつあるが、大規模言語モデル [1, 2, 9, 10] と同様に、しばしばハルシネーション（hallucination）[11, 12, 13] が生じる問題がある。

ハルシネーションとは、文生成時にモデルの入力情報や一般の社会事実に基づかない無関係な情報を生成してしまう現象であり、本研究では入力動画の内容とは無関係または事実ではない内容をモデルがキャプション文として生成してしまう事象を指す。

図 1　本データセットのアノテーション例。動画中の女性は”chair”に座り、背景にはピンクの”wall”が存在しているが、生成されたキャプション文はそれぞれ誤って”table”、”chair”と生成している。本データセットは、このような修正とそのハルシネーション部分のスパンをアノテーションしたものである。

このような事象は、動画キャプション生成技術の実用化を考える際に、ユーザーに誤った情報を提供する危険性を孕んでいる点で問題と言えよう。

動画キャプション生成におけるハルシネーションの研究には、Liu [14] らの研究がある。彼らは、約 6500 件の動画キャプション文に対してハルシネーション箇所をアノテーションし、分析を行った。また、Nasib [15] らは、動画キャプション生成のためのハルシネーションの評価指標 COAHA を提案し、動画中の物体や行動に関するハルシネーションの分析を行った。一方、Wang ら [16] や Zhang ら [17] は、動画質問応答タスクにおいて、新しいベンチマークを作成し、動画質問応答タスクのハルシネーションを抑制する方法を提案した。しかしながら、Liu [14]

らの研究では、モデルが「どこで」誤りが生じるかの誤りのスパンを明確にしたが、モデルが「どのように」誤るのかはアノテーションされていない。具体的には、”Dogs are running in the park.”というキャプション文に対し、”Dogs”がアノテーションされた場合に、映像中に犬が存在しないのか、複数匹存在しないのか、あるいは、動物種を誤認しているのかは明確ではない。そのため、誤りが「犬」と「別の動物」の種別間違いなのか、もしくは「1 匹」と「複数匹」の数的誤りなのかといった誤り方が分からない問題がある。このように、先行研究では、キャプション文内の誤りスパンに対して修正情報が付与されていないことで、モデルが「どのように」誤るのかを分析することが難しく、未だ取り組まれていない課題のひとつとなっている。

本研究では、このような課題の解決に向けて、動画キャプション文に対するハルシネーションが生じたスパンと共に内容修正をアノテーションした新しいデータセットを構築した。具体的には、複数の動画データセットから収集した 2,000 本の動画に対して、3 つのマルチモーダル LLM を用いて計 6,000 件のキャプション文を生成し、それらにアノテーションを行った。さらに、このデータセットを用いて動画キャプションにおけるマルチモーダル LLM のハルシネーション分析を行った。具体的には、修正前後の動画キャプション文の変化に注目し、単語の出現頻度や品詞といった言語的特性と、画質の明るさや画面の動きの大きさといった動画の視覚的特性の双方からマルチモーダル LLM に対するハルシネーションの分析を行った。

2 データセット

2.1 データセットの構成

本データセット[1)]は、動画データ、修正前のキャプション文 (以降、モデル生成文と表記)、修正後のキャプション文 (以降、モデル修正文と表記) の 3 点から構成される。動画データは、動画研究で一般的に用いられている MSR-VTT [18] と FAVDBench [19] の 2 種類の公開動画データセットから、それぞれ 1,000 件の動画データを選定し、合計で 2,000 件の動画データから構成される。次に、これらの動画データに対するキャプション文生成を、3 種類の VideoLLM を用いてそれぞれ行って (Sec。A.1)、モデル生成文を収集した。具体的には、Video-LLaMA [3]、VideoChat [4]、GPT-4v [20] の 3 種類のモデルに対して、前述の 2,000 件の動画データに対するモデル生成文を生成し、合計 6,000 件のモデル生成文を収集した。最後に、これらの 6,000 件のモデル生成文に対して、アノテーション作業 (Sec.A.2) として、誤りが含まれる場合にはそのスパンを特定することに加え、そのスパンの修正も併せて実施した。これにより、1,401 件のモデル修正文を収集した。図 1 は、ハルシネーションを含むデータとそのモデル生成文とモデル修正文を示した例である。モデル生成文にハルシネーションが含まれる場合、図 1 のようにそのスパンと修正内容が記録される。

1) 本研究で構築したデータセットは、公開に向けて準備を進めている。

表 1 動画データ 2000 件に対する各モデルによるキャプション文のハルシネーション有無のサンプル数を示す。

モデル	あり	なし	合計
Video-LLaMA	1410	590	2000
VideoChat	1423	577	2000
GPT-4v	1766	234	2000
合計	4599	1401	6000

2.2 データセットの性質

表 1 は、各 3 つのモデルに対してハルシネーションを含んだデータ数を示しており、全体で 1401 件の動画キャプション生成におけるハルシネーションを含んだサンプルが得られた。表 2 より、ハルシネーションを含むサンプルにおける平均誤り数は 1.22 個程度となっており、誤りスパン内の平均単語数は 2.86 単語である。

次に、アノテーション実施によるキャプション文の修正前後の CLIP 類似度と文字数の変化を比較した。CLIP 類似度は、動画とキャプション文がどの程度合っているかを示す指標として用いた。表 3 は修正前後のそれぞれの比較値を示しており、モデル生成文に比べモデル修正文の方が CLIP 類似度は 0.005 程度ほど僅かに上昇しているが、平均文字数は 12 文字ほど減少していることが分かる。これは、アノテーションによって言い換えや不要な誤りスパンが削除されたことで、キャプション文がより短い文章で正確に動画の内容が表現されていることを示している。

表 2　データセットの統計情報を示す。表は、スパン内に含まれる平均の単語数と、１サンプルあたりの平均スパンタグ数を示している。

指標	値
スパン内の平均単語数	2.86
スパンタグの平均数	1.22

表 3　表は、修正前後のキャプション文に対する CLIP 類似度、Perplexity、および平均文字数の比較を示している。Perplexity の計算には、GPT-2 モデルを用いた。

	平均 CLIP 類似度	平均 Perplexity	平均 文字数
修正前 (モデル生成文)	0.2935	36.79	84.39
修正後 (モデル修正文)	0.2984	107.49	72.66

3　ハルシネーションの分析

３つのマルチモーダル LLM によるモデル生成文とモデル修正文を比較し、言語情報と視覚情報の 2 つの観点からハルシネーション分析を行う。

3.1　言語情報に基づく分析

表 4 は、キャプション文の誤りスパンの正誤要素に含まれている単語の共起ペアを頻度順に並べたものを示している。共起ペアの頻度計算の際には、単語を見出し語化して行った。まず、表 4 の (a) では、修正前後で”person”と、”man”や”woman”がよく共起することから、モデルは人物表現の粒度が粗くなる可能性を示唆している。次に、(b) では、”room”と”stage”や”hall”がよく共起しており、機能や用途が異なる空間を表現する用語に対する誤りが生じやすいことを表している。同様に、(c) は動画”walk”、”sit”や”drive”と”stand”がよく共起し、オブジェクトの動きといった視覚的特徴に誤り[2)]が生じやすいことを示している。このような事例は、モデルが視覚的特徴に基づく用語に対する間違いがよく生じやすいことを示唆している。

次に、モデル生成文の品詞を分析する。表 5 は、修正前後のキャプション文の品詞[3)]の割合変化を示している。修正後では、名詞の比率が約 8%程度上昇している一方で、副詞とその他が減少している。これは、アノテーションによりハルシネーション部分の曖昧な表現や非事実の表現が削除された、または、より直接的に具体的な名詞を伴った表現に置き換えられたことを示唆している。一方、修正前後で動詞と形容詞の比率はほとんど変化しておらず、動画キャプションとして重要な「誰が何をするか」や「どんな状態であるか」を表す基本的な文構造は保存されていることが推察できる。これは、モデルは入力動画の主要な内容を反映した文を生成できるものの、動画内の対象や内容に関する詳細や適切な修飾表現の生成が未だ容易でない可能性を表している。

表 3 は、キャプションの修正前後の CLIP 類似度と GPT-2 モデルの Perplexity の値を示している。表 3 より、モデル修正文は CLIP 類似度がわずかに向上しており、動画との類似度が高まっていることが分かる。一方で、平均文字数の減少が示す通り単語の削除等による誤りの修正に反して、Perplexity の値は大きく増加しており、モデル修正文の文表現の流暢さ (自然さ) が大きく劣化していることが分かる。これは、モデル生成文が、入力動画の視覚情報に比べ、マルチモーダル LLM の言語生成機能を担う Decoder の影響を強く受けていること (詳細は Sec.A.3 を参照) が示唆される。

表 4　表は、キャプション文の誤りスパンに対する正誤要素に含まれている単語に対する共起ペアを頻度順に並べたものの一部を示す。表の中の”(empty)”は、誤りスパン内の単語が削除されたことを示す。

	修正前	出現数	修正後	出現数
(a)	person	42	(empty)	14
			man	10
			woman	10
(b)	room	45	stage	12
			(empty)	6
			hall	2
(c)	walk	12	stand	56
	sit	6		
	drive	5		

表 5　表は、修正前後における品詞の割合比較を示す。「その他」は、名詞、動詞、形容詞、副詞以外の品詞群を表す。

	修正前	修正後
名詞	74.84%	82.55%
動詞	4.47%	4.47%
形容詞	5.20%	5.15%
副詞	3.95%	1.57%
その他	11.54%	6.26%

2)　図 4 の (b) にサンプルを記載している。
3)　品詞推定には、nltk ライブラリ (v3.9.1) を用いた。

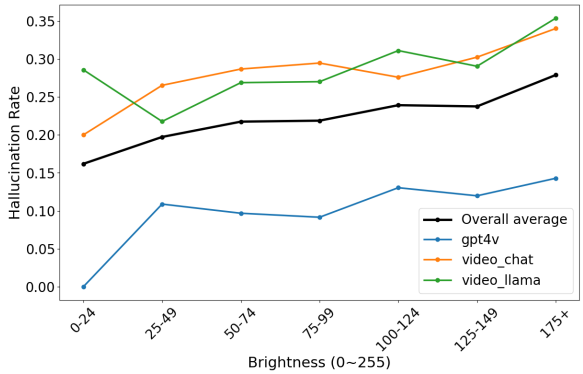

図 2　動画の明るさ毎のサンプル群に対するハルシネーションの発生率を示した図。縦軸がハルシネーション率を表し、横軸が動画の明るさを示している。動画の明るさが大きさに比例して、ハルシネーション率が増加していることが分かる。

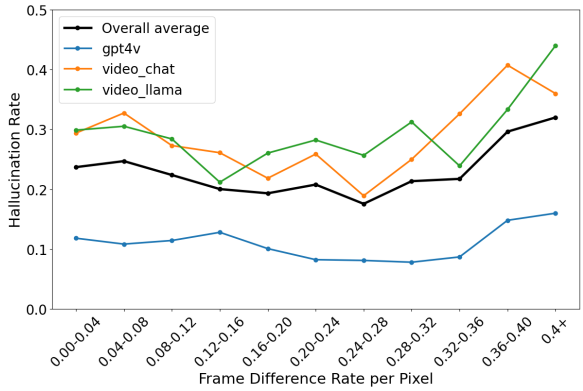

図 3　動画内の動きの激しさとハルシネーションの発生率を示した図。縦軸はハルシネーション率を表し、横軸は動画中の動きの激しさを示す。動画中の動きの激しさは、ピクセルあたりの frame difference を算出し、0〜1 区間のスコアとして用いる。

3.2　視覚情報に基づく分析

次に、入力動画の性質がキャプション文に与える影響に関する分析を行う。具体的には、人が動画を視聴する際に視覚情報として直感的に認知可能な「動画の明るさ」や「動きの激しさ」に注目し、ハルシネーションとの関係を分析する。

i) 動画の明るさ: 図 2 は、動画の明るさとハルシネーション発生率の関係を示しており、動画の明るくなるほどハルシネーション発生率が増加傾向にある。明るさが大きな動画を確認したところ、動画画面全体を白色が占めるような映像や広告の動画が多く、シーン構成や演出の視覚的理解が比較的難しい傾向にあることが分かった。このような動画の性質は、モデルのハルシネーションを引き起こす要因となりうる。その一例として、図 4 の (a) を示す。図

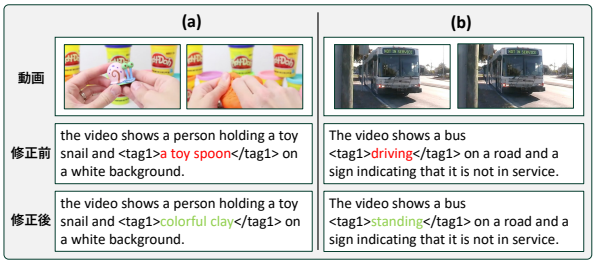

図 4　ハルシネーションを含むキャプション例。

4 の (a) は、色付き粘土の商品広告動画であり、完成品の紹介や粘土の使用方法に関する説明など、複数のシーンから構成されている。この動画に対するモデル生成文では、”colorful clay”が正しい表現に対して、”a toy spoon”と誤生成していることが分かる。これは、広告動画のように複数のシーンの移り変わりが頻繁なコンテンツに対して、モデルの視覚的特徴の抽出性能が正確ではないことを示唆している。

ii) 動きの激しさ: 図 3 は、動画中の動きの激しさとハルシネーション発生率の関係を示しており、画面内の動きが小さい場合と大きな場合でハルシネーション発生率が増加傾向にある。それらの動画内容を調べたところ、動きが小さい動画は広告や人工的な映像を含むものが多く、動きが大きな動画ではシーンの移り変わりが頻繁なものが多い傾向にあった。前述同様にシーンの移り変わりが頻繁な動画内容に対しては一貫性のあるキャプション文生成が難しく、誤りが生じやすいと推察できる。一方で、動きの小さな動画においては、図 4 の (b) のような場合が見られた。図 4 の (b) は、静止しているバスを映した動画であるが、モデル生成文には”driving”という動詞が生成されている。これは、動画中の乗り物は一般的に動いている様を映すといった視覚モデル (encoder) 側のバイアスがキャプション文に影響を与えている可能性があり、モデルが静止と動作の状態の区別が難しいことが示唆される。

4　おわりに

本研究は、動画キャプション生成におけるマルチモーダル LLM のハルシネーション分析とデータセット構築を行った。分析結果から、ハルシネーションの内容が入力動画の視覚情報よりマルチモーダル LLM の Decoder の影響を強く受けることや動画の明るさや動きの視覚情報と関連する傾向にあることが分かった。今後、モデル別のさらなる分析とこれらの知見に基づいてハルシネーションを抑制するためのモデル設計や手法の研究を予定している。

謝辞

この研究の遂行にあたり、多大なる貢献をいただきました西村太一氏に感謝いたします。特に、データセットの構築における彼の貢献は、本研究の基盤を支える重要な役割を果たしました。

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[3] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In **Conference on Empirical Methods in Natural Language Processing**, 2023.

[4] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. **arXiv preprint arXiv:2305.06355**, 2023.

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **Advances in neural information processing systems**, 2024.

[6] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In **AAAI Conference on Artificial Intelligence**, 2024.

[7] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In **Conference on Empirical Methods in Natural Language Processing**, 2023.

[8] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. **Conference on Empirical Methods in Natural Language Processing**, 2024.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. **Journal of Machine Learning Research**, 2023.

[10] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, 2022.

[11] Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. An audit on the perspectives and challenges of hallucinations in nlp. In **Conference on Empirical Methods in Natural Language Processing**, 2024.

[12] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, 2023.

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Computing Surveys**, 2023.

[14] Hui Liu and Xiaojun Wan. Models see hallucinations: Evaluating the factuality in video captioning. In **Conference on Empirical Methods in Natural Language Processing**, 2023.

[15] Nasib Ullah and Partha Pratim Mohanta. Thinking hallucination for video captioning. In **Asian Conference on Computer Vision**, 2022.

[16] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. **arXiv preprint arXiv:2406.16338**, 2024.

[17] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. **arXiv preprint arXiv:2409.16597**, 2024.

[18] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In **Conference on Computer Vision and Pattern Recognition**, 2016.

[19] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, et al. Fine-grained audible video description. In **Conference on Computer Vision and Pattern Recognition**, 2023.

[20] OpenAI. Gpt-4v(ision) system card. 2023.

A Appedix

A.1 動画キャプションの生成操作

動画キャプション生成の手続きを説明する。本研究で用いる 3 つのマルチモーダル LLM に対して、動画とプロンプト文を入力して与え、動画キャプション文を出力を行う。Video-LLaMA と VideoChat のモデルには、プロンプト文として”Describe this video for details.”という文を用いて動画キャプションの生成を行った。また、GPT-4v では、OpenAI の API に動画データを直接入力することはできないため、動画中からフレーム画像を 5 枚をランダムサンプリングし、それらを時系列に並べた画像リストを入力として用いることで、キャプション文の生成を行った。

A.2 アノテーションの仕様

本データセットは、人手により各モデルのモデル生成文に対してハルシネーションが含まれるか否かの確認を行い、ハルシネーションが含まれる場合にはそのスパンと修正を行うアノテーション作業を実施することでデータセットの構築を行った。具体的には、ハルシネーションの定義を“動画の内容からは判別不能である、または、動画内容と矛盾する”こと、として定め、動画データとモデル生成文のペアに対して、モデル生成文にハルシネーションが含まれるかを人手確認した。次に、ハルシネーションが含まれる場合にはそのスパンを定めるため、モデル生成文とモデル修正文のそれぞれの誤り対応箇所に "<tag>〜<\tag>" というタグで囲み、以下のアノテーション規則に沿って修正文の作成を実施した。

- できる限り修正箇所の範囲が小さくなる（狭くなるように）誤りの修正を行う。
- 修正操作は、言い換えによる”訂正操作”と誤り箇所の除去を目的とする”削除操作”の 2 つの手続きで行う。
- ひとつの説明文に対して、複数の修正箇所が検出された場合は検出箇所すべてを修正する。（削除操作の場合は、検出箇所のみを記録する）
- 説明文全体が誤っており修正が困難な場合は、「修正が困難」というラベルを説明文に付与する。

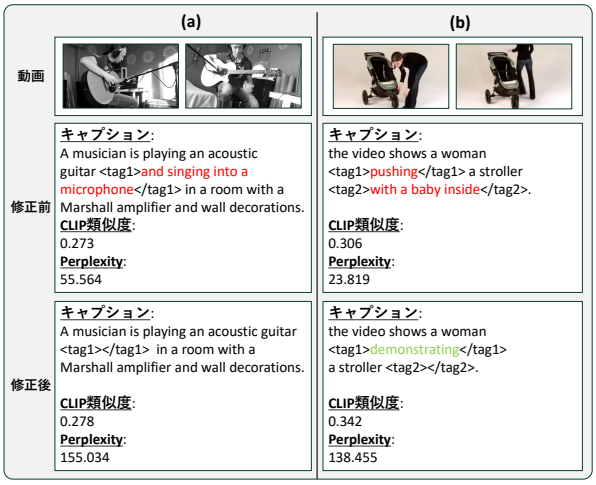

図 5 図は、ハルシネーションを含むキャプションの事例を示している。いずれも動画中に存在しない動きや物体について不必要な情報が付加されている。

A.3 LLM の言語生成機能の影響を受けるハルシネーションに対する考察

図 5 の 2 例は、どちらも修正後のキャプション文に削除操作が適用されたものを示している。(a) はアコースティックギターを演奏している男性の動画で、モデルによる生成キャプション文に”and singing into a microphone”とあり、実際には行われていない歌唱行為が誤って付与されている。(b) はベビーカーを紹介する女性の広告風の動画で、モデルによる生成キャプション文には”with a baby inside”と生成され、実際には存在しない赤ちゃんの表現が付加されており、視覚情報を無視して生成がなされている。どちらも修正後のキャプション文は削除操作にて正確な動画内容を反映した文へと変更がなされ、CLIP 類似度のスコアも高くなっていることから、動画キャプション文の正確さは高まったとみなしてよいだろう。

しかしながら、GPT-2 による Perplexity の評価を行った場合には、どちらの修正後キャプション文も誤った内容を含む修正前に比べ、スコアが大きくなっており、文の尤もらしさは低下したことを示している。これは、(a) の”ギター”と”歌唱”や、(b) の”ベビーカー”と”赤ちゃん”といった一般的に共起しやすい言語表現をマルチモーダル LLM の言語モデル (Decoder) 側が学習し、それが強く影響することでキャプション文の誤りが生成される可能性を示唆している。すなわち、動画内容に基づく表現ではなく文章表現上 (生成文字の順序上) の尤もらしさを優先した結果として、誤ったキャプション文の生成に繋がっているのではないかと筆者らは考えている。