C2-01

長尺動画生成タスクにおける メタ評価ベンチマーク

松田 陵佑[1], 工藤 慧音[1,2], 吉田 遥音[1], 清水 伸幸[3], 鈴木 潤[1,2,4]
([1] 東北大学, [2] 理化学研究所, [3] LINEヤフー株式会社, [4] 国立情報科学研究所 LLMC)
is-failab-research@grp.tohoku.ac.jp

概要

- 長尺動画生成モデルの評価システムの性能を メタ評価するための**SLVMEval ベンチマーク**を提案
- **人工的**に**平均19分 最大約3時間**の 高品質/低品質の**長尺動画ペア**を作成

! 既存の評価システムは9観点で**人間性能に及ばない** **特にプロンプトと動画の一貫性**に関して性能が低い

背景

動画生成(T2V)モデルの性能は進化しているが
長尺動画生成(T2LV)と**その評価手法は未開拓**

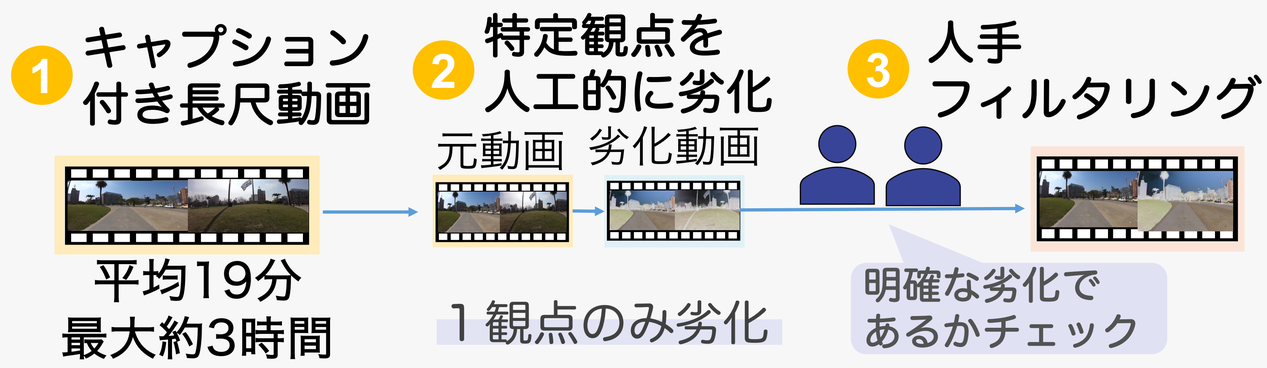
? **正しく測定できないものは改善出来ないのでは？**
- 既存のT2V評価モデルは数秒の動画で 学習された評価モデル [He+'2024]
- 人間が評価するのはコスト・時間的に困難

! **評価モデルをメタ評価するベンチマークの整備が必要**

データセット

SLVMEval の概要

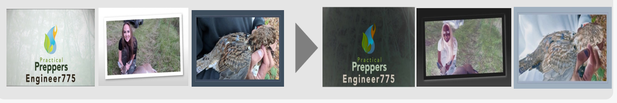
現在のT2Vモデルの**生成時間を大きく超える&**
観点別の精緻な評価を可能にしたい！
→ **長尺動画に評価観点別に**人工的に劣化処理を施す



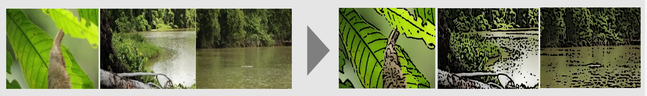
劣化処理と動画例

見た目の品質

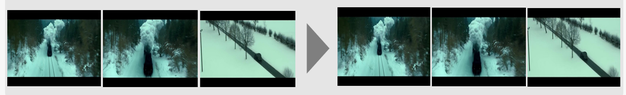
美的品質:コントラストを低下



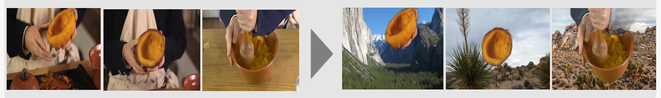
スタイル:不自然なスタイル変換



技術的品質:解像度を低下



背景の一貫性:無関係背景に置換

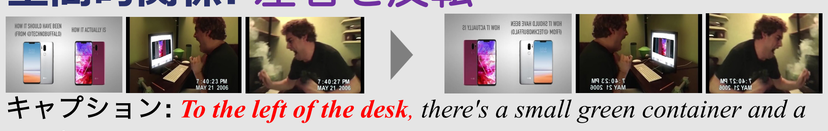


プロンプトと動画の一貫性

物体の完全性: 指定の物体削除

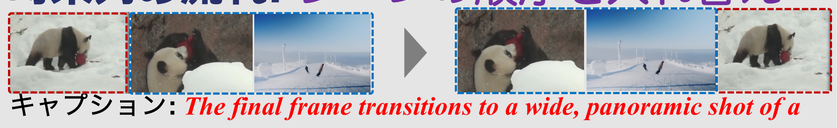
キャプション：*The subject, dressed in a dark suit with a white shirt, is positioned against a blurred.*

空間的関係: 左右を反転



キャプション: *To the left of the desk, there's a small green container and a set of gray.*

時系列の流れ: シーンの順序を入れ替え



キャプション: *The final frame transitions to a wide, panoramic shot of a snow-covered landscape.*

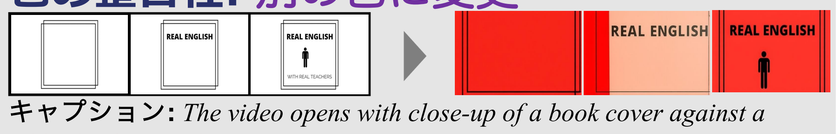
動きの度合い: 静止画に置換

キャプション: *There is a rapid change to a blurry image of a vintage car-a red.*

網羅性: シーンを削除

キャプション: *They walk from one side of the frame to the other, crossing a small yard with a traditional.*

色の整合性: 別の色に変更



キャプション: *The video opens with close-up of a book cover against a white background.*

実験設定

設定1.ペアワイズ評価

元動画 vs. 劣化動画 のペアのどちらが高品質かを
評価システムに**評価**させて正答率を測る

Q. どちらの動画がある観点において高品質か?

設定2.評価システム

- **VLM-as-a-judge** (GPT-5, Qwen3-VL)
  - **動画ベース評価:** VLMに２つの動画を入力し評価
  - **テキストベース評価:** VLMに事前にキャプションを 作成させて２つのキャプションを入力し評価
- **CLIPScore** [Hessel+'2021]
- **VideoScore** [He+'2024]

実験結果

結果1.全体の分析

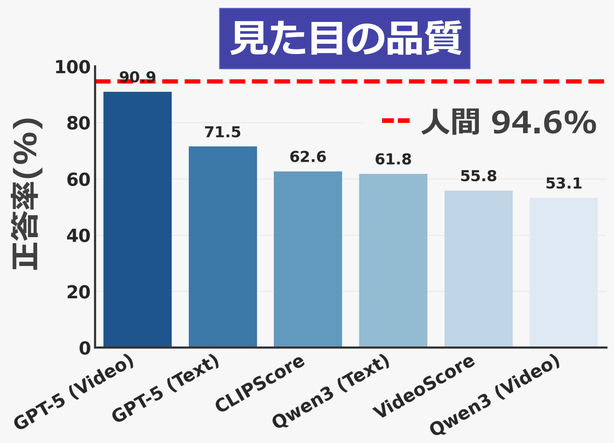
人間は90%以上の正答率

! **GPT-5でも9観点で 人間に及ばない**

→**既存のVLM等は簡単な タスクでさえ解けない**

特に**プロンプトと動画の 一貫性**は性能が低い

見た目の品質



プロンプトと動画の一貫性

結果2.動画長への脆弱性

人間は**頑健性が高い**

! 評価システムは複数観点で
動画長の増加に伴い正解率が低下の傾向