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] 国立情報学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

動画生成モデルの評価システム自体の性能をメタ評価するためのベンチマーク SLVMEval を提案する．SLVMEval は合成的に構築された，高品質動画と低品質動画のペアからなる．これらの動画は平均約 19 分，最大約 3 時間の長尺動画となっている．評価システムは，与えられた動画ペアのどちらが高品質であるかを識別できる割合（正解率）によって評価される．実験の結果，既存の自動評価システムは 10 観点中 9 観点で人間の正解率に及ばず，特にプロンプトと動画の一貫性に関する観点で性能が著しく低いことを明らかにした．

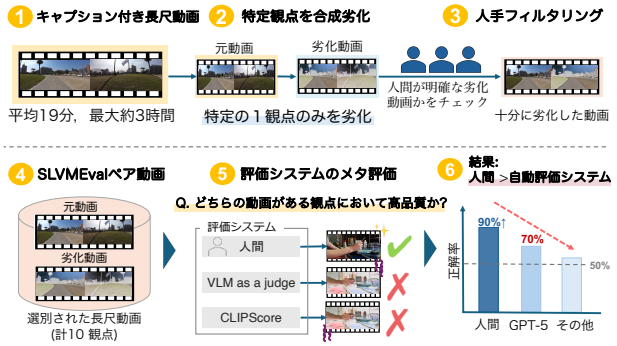

図 1　SLVMEval の概要．長尺の元動画に対し，特定の観点に基づいた劣化処理を施すことで動画ペアを構築する．その後，人手フィルタリングを経たデータセットを用いて，既存の自動評価システムのメタ評価を検証する．

1　はじめに

映画やテレビドラマなど，私たちが日常的に消費する動画コンテンツの多くは，数十分から数時間に及ぶ長尺なものである．一方で，近年の**テキストからの動画生成（T2V）**モデルの進展 [1, 2] は目覚ましいものの，その出力は依然として数秒から数十秒程度に限られている．これに対し，理論上任意の長さの動画を生成可能な T2V モデル [3, 4] も登場しており，**テキストからの長尺動画生成（T2LV）**モデルは次なるフロンティアとして注目されている．

T2LV モデルの研究開発を進める上で，その性能を正しく測る自動評価システムの確立は不可欠である．既存研究では，VideoScore [5] などの自動評価指標が提案されている．しかし，これらは最大数十秒の短尺動画向けに設計されており，長尺動画に対する評価能力は未知数である．また，評価システム自体の性能を測るためのメタ評価用データセットも公開されている（VBench [6]，UVE [7]）．これらについても短尺動画のみを対象としており，文脈理解や時間的一貫性が重要となる長尺動画特有の観点についてメタ評価を行うことは困難である．

本研究では，T2LV の評価システムをメタ評価するための新たなベンチマーク **SLVMEval** を提案する (図 1)．本ベンチマークの目的は，評価システムが長尺動画を評価する上で最低限必要な能力，すなわち「人間であれば容易に判断できる長尺動画の品質差を正しく識別できるか」を検証することである．既存の長尺動画データセットに対し，特定の観点に対応した劣化処理をすることで高品質動画（元動画）と低品質動画（劣化動画）のペアを作成する．これらの動画時間は数分から数時間にわたり，現在の T2V モデルの生成範囲を大きく超える．

実験では，SLVMEval を用いて既存の自動評価システムおよび人間のアノテータを対象にメタ評価を実施した．その結果，人間のアノテータでは全ての観点で 84%以上の高い正解率で評価を行うことができた．それに対して，既存の自動評価システムは 10 観点中 9 観点で人間の正解率に及ばなかった．特に，動画とテキストの意味的・時間的な一貫性を問う観点での性能は低く，さらに動画が長くなるほど正解率が低下する傾向も確認された．また，人手フィルタリングを行った場合と行わなかった場合での評価結果に高い相関が見られたことから，コストのかかる人手フィルタリングなしでも信頼性の高いベンチマーク拡張が可能であることも明らかにした．

2　メタ評価の枠組み

本研究では，VBench [6] や UVE [7] と同様に与えられた 2 つの動画のうちどちらがより高品質かを選択するペアワイズ比較による枠組みにおける評価システムの性能のメタ評価を実施する．

評価システム　あるプロンプト p に対し，2 つの T2V モデルが生成した動画ペア $\{u_p, v_p\}$ が与えられる．評価システム e はこの動画ペアとプロンプト p を入力として，より質が高いと思われる動画 (u_p または v_p) を選択する．

メタ評価指標　システムの評価性能を測る指標として，正解率を用いる．評価データセット $\mathcal{D} = \{(p_i, \{v^+_{p_i}, v^-_{p_i}\})\}_{i=1}^N$ は，複数の高品質動画 $v^+_{p_i}$ と低品質動画 $v^-_{p_i}$ のペアからなる．評価データセットの全ての動画ペアに対して評価を行った時，評価システムが高品質動画 v^+_p を選択した割合が高いほど，そのシステムは T2LV タスクの自動評価指標として高性能であると考えられる．

3　データセット

本研究の目的である，評価システムが最低限の能力を満たしているかを測るため，特定の観点のみが異なる動画ペアを人工的に作成する．元のデータセットとして，長尺かつ高密度なキャプションが付与されている Vript [8] を利用する．データセットに含まれる各動画は意味的にまとまりのある区間に分割されており，各区間に対して個別のキャプションが付与されている．これらを連結したものを動画全体のプロンプト p として使用する．

3.1　劣化処理

長尺動画の失敗の傾向を網羅することは困難であるため，包括的な短尺動画ベンチマークである VBench [6] を参考に 10 の評価観点を定義し，**見た目の品質**と**プロンプトと動画の一貫性**の 2 つのカテゴリに分類する．各観点 a について，元動画 v^+_p の一部のクリップに劣化処理を適用し，低品質動画 v^-_p を得る．以下では，各観点ごとの劣化処理の概要を述べる．[1)]

見た目の品質　視覚的な品質を重視して評価する 4 つの観点である．

- **美的品質**：コントラストを低下させ，美的品質を劣化させる．
- **技術的品質**：解像度を下げた荒いフレームに置き換える．
- **スタイル**：OpenCV [9] を用いて特定のスタイルに変換し，不自然なスタイル変化を作り出す．
- **背景の一貫性**：rembg [10] を用いて背景を除去し，ランダムな風景画像に置き換えることで，背景が一貫しない状況を作り出す．

プロンプトと動画の一貫性　動画の意味的内容を重視して評価する 6 つの観点である．

- **時系列の流れ**：連続するクリップの順序をランダムに入れ替え，イベントの順序を乱す．
- **網羅性**：一部のクリップを削除し，プロンプトで記述された内容が動画に含まれない状況を作る．
- **物体の完全性**：動画内に特定の物体を画像編集モデルを用いて消去する．
- **空間的関係**：特定の物体の左右位置についてプロンプトで言及がある区間の全てのフレームを左右反転させる．
- **動きの度合い**：動きに関する記述がキャプションに存在する区間の全てのフレームを，単一のフレームの繰り返しに置き換える．
- **色の整合性**：動画中の特定の物体の色を画像編集モデルを用いて他の色に変更する．

3.2　人手フィルタリング

劣化処理が対象の観点に則して意図した通り行われていることを保証するため，クラウドソーシングを用いて計 736 人によるフィルタリングを行った．各動画ペアについて 5 人のワーカーが劣化の成功度を A（明確に成功），B（部分的に成功），C（完全に失敗）の 3 段階で評価した．「C 評価が 1 つもない」かつ「A 評価が B 評価より多い」サンプルのみを最終的な評価データセットとして採用した．

3.3　データセットの統計量

フィルタリング後の評価データセットは 3,932 件の動画ペアで構成される．動画の長さは平均 1,141 秒 (≃ 19 分)，最長で 10,486 秒 (≃ 3 時間) である．[2)] また，15 の多様なドメインの動画から構成されている．表 1 は既存のデータセットとの統計量の比較を示しており，SLVMEval は長尺動画とプロンプトを含むことがわかる．

1)　劣化処理の詳細は付録 A を参照のこと．

2)　詳細な動画時間の分布は付録 A に記載．

表 1　既存ベンチマークと本ベンチマークの統計．VBench-Long [11] は自動評価のフレームワークのみを提供しており，人手アノテーションは含まない．また生成動画の長さは約 1 分を想定している [12]．

	人手評価	観点数	動画数	プロンプト数	最大動画時間 (秒)	平均プロンプト長 (chars)
UVE-Bench [7]	✓	15	1,045	293	6.1	73.68
VBench [6]	✓	16	21,110	968	3.3	41.32
VBench Long [11]		16	N/A	944	N/A	41.00
SLVMEval (**ours**)	✓	10	3,932	1,461	10,486.0	57,883.52

表 2　SLVMEval における各ベースラインシステムの正解率（%）．数値は正解率 % ± 95 % 信頼区間（パーセントポイント）を表す．青色太字は各観点における最高値，緑色は 2 番目の値を示す．チャンスレートは 50%である．

	見た目の品質				プロンプトと動画の一貫性					
	美的品質	技術的品質	スタイル	背景の一貫性	物体の完全性	色の整合性	動きの度合い	網羅性	空間的関係	時系列の流れ
動画ベース										
GPT-5	90.1±2.5	85.8±4.2	88.9±2.5	**98.9**±0.8	72.0±6.2	84.3±3.5	35.3±3.6	51.3±4.5	59.7±4.4	50.3±4.1
GPT-5-mini	84.0±3.0	48.1±6.1	78.0±3.2	95.2±1.6	66.5±6.5	69.4±4.5	31.5±3.5	45.7±4.5	51.1±4.5	43.7±4.1
Qwen3	55.7±4.1	51.9±6.1	55.3±3.9	49.7±3.7	38.5±6.7	48.8±4.9	50.0±3.8	52.6±4.5	51.7±4.5	50.2±4.1
テキストベース										
GPT-5	74.8±3.6	46.2±6.1	81.1±3.1	83.8±2.7	68.0±6.5	68.9±4.5	43.1±3.8	50.6±4.5	47.0±4.5	43.5±4.1
GPT-5-mini	75.0±3.6	53.8±6.1	79.6±3.2	81.1±2.9	65.5±6.6	71.8±4.4	43.8±3.8	50.6±4.5	51.1±4.5	41.2±4.0
Qwen3	51.6±4.1	50.0±6.1	72.4±3.5	73.0±3.3	51.0±6.9	61.0±4.7	52.7±3.8	51.7±4.5	50.2±4.5	55.6±4.1
CLIPScore	56.4±5.8	72.3±7.7	53.2±5.5	68.6±4.8	76.0±8.4	66.2±6.5	51.7±5.4	57.4±6.3	55.1±6.3	50.5±5.8
VideoScore	52.5±5.8	33.8±8.1	65.7±5.3	71.2±4.7	66.0±9.3	33.8±6.5	48.6±5.4	34.5±6.1	49.6±6.4	46.3±5.8
人間	**96.5**±2.1	**91.8**±4.7	**95.2**±2.4	95.0±2.3	**86.6**±6.7	**96.8**±2.4	**95.9**±2.1	**84.7**±4.6	**88.2**±4.1	**86.6**±4.0

4　実験設定

複数の既存の評価システムについて，SLVMEval における性能評価を実施する．評価指標には 2 章で述べた正解率を用いる．また，比較対象として人間のアノテータが評価した場合の正解率も測定する．

4.1　ベースライン評価システム

以下の評価システムを検証対象とした．

VLM-as-a-judge　2 つの設定で評価する．

- **動画ベース評価**：VLM に 2 つの動画を入力し，どちらの動画が高品質であるかの予測結果をテキストで出力する．
- **テキストベース評価**：前処理として，それぞれの動画を独立に VLM に入力しそのキャプションを得る．その後，2 つの生成したキャプションとプロンプトを VLM に与え，どちらのキャプションがプロンプトと整合しているかを予測する．

なお，提示順序バイアスを軽減するため，入力動画及びキャプションの順序を入れ替えた場合両方で評価を行う．VLM として GPT-5，GPT-5-mini [13]，Qwen3-VL-235B-Thinking [14, 15, 16] を利用する．

CLIPScore　動画内のフレームとプロンプト間の CLIPScore [17] を計算し，スコアが高い方の動画を選択する．モデルは長系列を扱うことが可能な Jina CLIP v2 [18] を使用する．

VideoScore　VideoScore-v1.1 [5] を用いて各動画のスコアを算出し，高い方を選択する．

5　結果と考察

主な結果を表 2 に示す．

ベンチマークの妥当性　人間の評価者は全 10 観点において 84.7%–96.8% という高い正解率を達成した．これは SLVMEval における高品質・低品質の差が人間にとって明確であり，評価システムに最低限求められる能力を測るベンチマークとして妥当であることを示している．

全体的な傾向　既存の評価システムは，10 観点中 9 観点で人間の正解率を下回った．特に「見た目の品質」カテゴリでは GPT-5 (動画ベース) が 85%以上の正解率を示したが，「プロンプトと動画の一貫性」カテゴリでは人間との差が顕著であった．例えば「時系列の流れ」や「動きの度合い」といった，前後の文脈理解が必要な観点では，多くのシステムがチャンスレートに近い性能にとどまった．これは，最新の大規模言語モデルであっても，長尺動画における意味的・時間的整合性の理解には限界があることを示唆している．

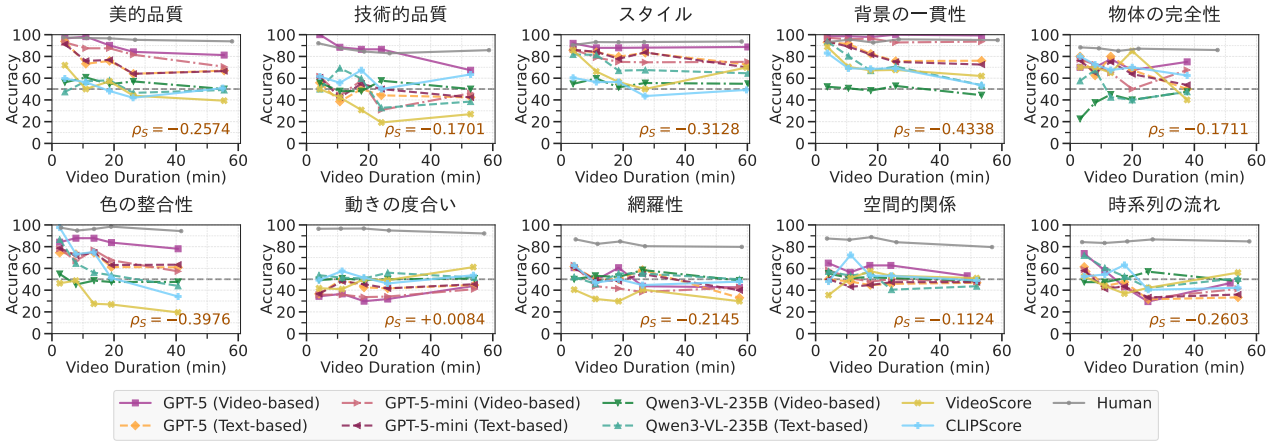

図 2　動画長と正解率の関係．データを動画長で 4 つのビンに分割し，各ビンでの正解率をプロットした．また，各観点・評価システムについて動画長と正解率のスピアマン順位相関（ρ_S）を算出した．

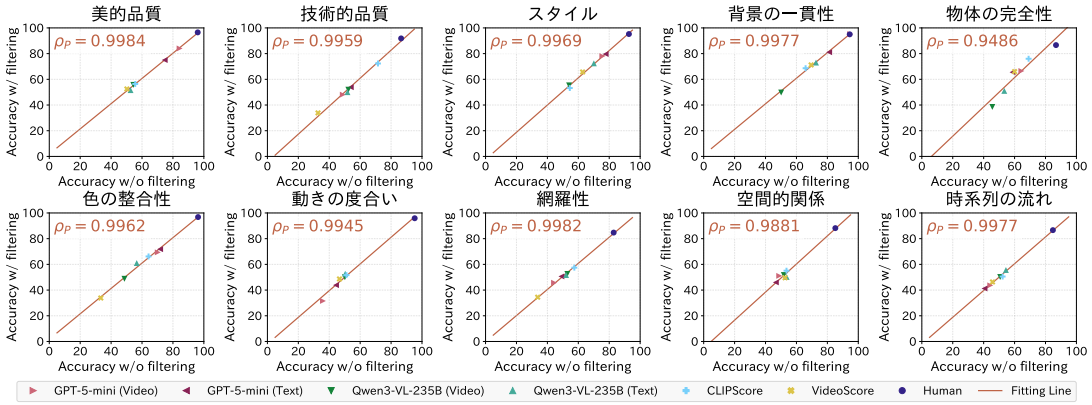

図 3　人手フィルタリング前後のデータを用いた正解率の関係．各観点において，フィルタリングなしのデータでの正解率（横軸）とありのデータでの正解率（縦軸）をプロットし，各観点毎にピアソン相関（ρ_P）を算出した．

CLIPScore の傾向　「物体の完全性」や「網羅性」で人間に次ぐ性能を示した一方で，「時系列の流れ」の観点での性能は低かった．CLIPScore は，フレームごとに独立に評価するため，時間構造の考慮が不得意であることに起因すると考えられる．

テキストベース評価の傾向　Qwen3 において，動画ベース評価と比較して「背景の一貫性」で 23.3 ポイント，「スタイル」で 17.1 ポイントの大幅な性能向上を示した．これらの観点では，言語情報の方が文脈を捉えやすい可能性を示唆している．

動画長に対する脆弱性　図 2 に動画長と正解率の関係を示す．「背景の一貫性」や「色の整合性」をはじめとする多くの観点で，人間の正解率は動画長に関わらず安定しているのに対し，自動評価システムは動画が長くなるにつれて正解率が低下する傾向が見られた．これは，既存システムが長い入力系列を処理する能力に限界がある可能性を示唆している．

データフィルタリングの必要性　図 3 に，3.2 節で述べた人手フィルタリングを実施した場合としなかった場合での，各評価システムにおける正解率の相関を示す．両者の間には極めて高い相関（$\rho_P > 0.94$）が見られた．これは，本研究で提案する劣化処理が十分に信頼性が高く，コストのかかる人手フィルタリングを行わずとも，ベンチマークを拡張可能であることを示唆している．

6　おわりに

本研究では，T2LV の評価システムをメタ評価するためのベンチマーク SLVMEval を提案した．既存の評価システムを用いた実験の結果，人間にとっては容易なタスクであっても，自動評価システムは多くの観点で十分な性能を発揮できないことが明らかになった．特に，動画とテキストの一貫性や，動画が長くなることによる性能低下に課題があることが示された．まず，T2LV 評価システムが SLVMEval で示された欠点を改善することが，今後の T2LV モデルの開発を効率的に進めるための一つの目標になると期待される．

謝辞

研究遂行にあたりご助言ご協力を賜りました Tohoku NLP グループの皆様に感謝申し上げます．また，Yahoo!クラウドソーシングを通してアノテーションに参加していただいたワーカーの皆様にも感謝申し上げます．本研究は，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research)，および，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」，JSPS 科研費 JP25KJ0615 の助成を受けたものです．本研究成果の一部は，九州大学情報基盤研究開発センター研究用計算機システムの「一般利用」を利用して得られたものです．

参考文献

[1] Google DeepMind. Veo: a text-to-video generation system. Technical Report Veo 3 Tech Report, Google DeepMind, 2025. Accessed: 2025-07-26.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024.

[3] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In **International Conference on Learning Representations**, 2023.

[4] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025.

[5] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2105–2123, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[6] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, 2024.

[7] Yuanxin Liu, Rui Zhu, Shuhuai Ren, Jiacong Wang, Haoyuan Guo, Xu Sun, and Lu Jiang. Uve: Are mllms unified evaluators for ai-generated videos?, 2025.

[8] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, **Advances in Neural Information Processing Systems**, Vol. 37, pp. 57240–57261. Curran Associates, Inc., 2024.

[9] G. Bradski. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, 2000.

[10] Gatis, Daniel and contributors. rembg: Image background removal tool, 2022. GitHub repository.

[11] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024.

[12] Ziqi Huang. Comprehensive benchmark suite for video generative models: Vbench. CVPR 2024 Workshop slides, June 2024. Accessed: 2025-11-12.

[13] OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025. Accessed: 2025-11-12.

[14] Qwen Team. Qwen3 technical report, 2025.

[15] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. **arXiv preprint arXiv:2502.13923**, 2025.

[16] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. **arXiv preprint arXiv:2308.12966**, 2023.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[18] Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images, 2025.

[19] mertcobanov and contributors. Nature Dataset: Landscape background images, 2024. Hugging Face dataset; ID: mertcobanov/nature-dataset.

[20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In **Computer Vision – ECCV 2024**, pp. 38–55, Cham, 2025. Springer Nature Switzerland.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 10684–10695, June 2022.

[22] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

A　付録

図 4　SLVMEval における元動画と劣化動画のペアの例．10 の各観点について，元動画（左）と特定の劣化を施した動画（右）を示している．上段は「見た目の品質」カテゴリ，下段は「プロンプトと動画の一貫性」カテゴリに対応する．

Algorithm 1 Overview of Φ_a^{low}

Require: Prompt $p \in \mathcal{P}$, Video $v_p \in \mathcal{V}$
Ensure: v_p^-

$S \leftarrow \textsc{SampleIdx}(M_{v_p}, 5)$　▷ 5 つのクリップをランダムサンプリング
$F \leftarrow [\,]$
for $m \leftarrow 0$ **to** $M_{v_p} - 1$ **do**
　$clip \leftarrow c_{v_p,m}$
　if $m \in S$ **then**
　　$F \leftarrow F \cup \textsc{DegradeClip}(clip,\ p_m)$
　else
　　$F \leftarrow F \cup clip$
　end if
end for
$v_p^- \leftarrow \textsc{Concat}(F)$
return v_p^-

劣化処理アルゴリズム　3.1 節で定義した，各観点ごとの劣化動画生成プロセス Φ_a^{low} を Algorithm 1 に示す．Φ_a^{low} では，動画全体から特定のクリップ（本実験では 5 つ）をサンプリングし，それらに対してのみ観点固有の劣化関数 DegradeClip を適用することで，局所的な劣化を含む長尺動画を作成した．

劣化処理の詳細　いくつかの観点における DegradeClip の具体的な処理内容は以下の通りである．

【見た目の品質】美的品質：FFmpeg の eq フィルタを用い，対象クリップのコントラスト係数を −0.8 に設定する．輝度成分を反転させた後，ダイナミックレンジを 80%に圧縮することで，視覚的な魅力を損なわせる．**技術的品質**：動画フレームの長辺を 512px に統一した後，対象クリップのみ LANCZOS 法を用いて長辺 256px にダウンサンプリングし，直後に 512px へアップサンプリングすることで，解像度不足によるノイズを再現する．**スタイル**：OpenCV [9] を用い，アニメ調，詳細強調，油絵，色鉛筆，水彩画の 5 つのスタイルからランダムに 1 つを選択して適用する．**背景の一貫性**：Qwen3-8B [14] によりキャプションに背景情報が含まれるクリップを抽出する．rembg [10] を用いて被写体の背景を除去した後，nature-dataset [19] からランダムにサンプリングした画像を合成し，フレーム間の背景の不整合を作り出す．**【プロンプトと動画の一貫性】物体の完全性**：Qwen3-8B を用いてキャプションから主要な物体名を抽出し，Grounding DINO [20] で対象物体のバウンディングボックスを検出する．その後，Stable Diffusion Inpainting [21] を用いてその物体を映像から消去する．**色の整合性**：色情報（例:「赤い車」）を含む物体を特定し，Qwen-Image-Edit [22] を用いて，対象物体の色のみを記述とは異なる色（例：赤 → 青）に変更する．

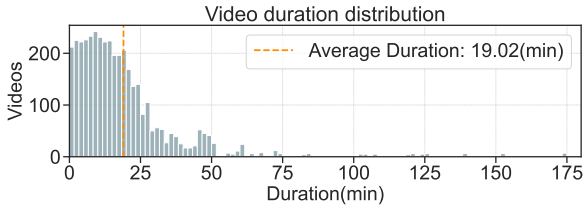

図 5　SLVMEval における動画時間の分布．100 秒刻みのビンで集計している．

劣化動画の例　図 4 に，SLVMEval における各観点の元動画と劣化動画のペアの例を示す．

動画時間の分布　構築した SLVMEval データセットに含まれる動画時間の分布を図 5 に示す．

アノテーション　本研究では Yahoo!クラウドソーシングを用いて人間のアノテータを募集した．最終的に計 3,793 タスクを計 736 人のワーカーにより実施した．報酬単価は日本の平均最低賃金を上回る実質時給となるよう設定しており，総支払額は 408,520 円であった．

またデータ品質を高めるため，回答時間が極端に短い・回答の合意率が低い・タスク正解率の低い等の基準に基づき信頼性の低いアノテータは以降のタスク割り当てから除外した．