VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation

[1,2,†] Xuan He*, [1,†] Dongfu Jiang*, [1,3] Ge Zhang, [1] Max Ku,
[1]Achint Soni, [1]Sherman Siu, [1]Haonan Chen, [1]Abhranil Chandra, [1]Ziyan Jiang, [1]Aaran Arulraj, [4]Kai Wang, [1]Quy Duc Do, [1]
Yuansheng Ni, [2]Bohan Lyu, [1]Yaswanth Narsupalli, [1]Rongqi Fan, [1]Zhiheng Lyu, [5]Bill Yuchen Lin, [1,†] Wenhu Chen

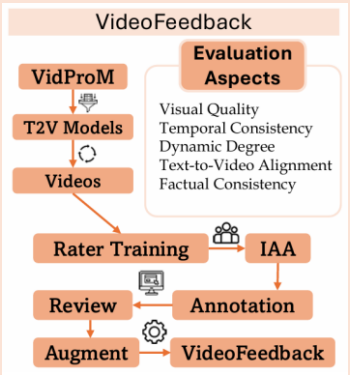
*Equal Contribution

[1]University of Waterloo, [2]Tsinghua University, [3]StarDust.AI, [4]University of Toronto, [5]AI2

Introduction

None of the existing metric is able to provide reliable scores over generated videos. The main barrier is the lack of large-scale human-annotated dataset.
In this paper, we release VideoFeedback, the first large-scale dataset containing human-provided multi-aspect score over 37.6K synthesized videos from 11 existing video generative models.
We train VideoScore based on VideoFeedback to enable automatic video quality assessment. Experiments show that the Spearman correlation between VideoScore and humans can reach 77.1 on VideoFeedback-test, beating the prior best metrics by about 50 points. Further result on other held-out EvalCrafter, GenAI-Bench, and VBench show that VideoScore has much higher correlation with human judges than other metrics.

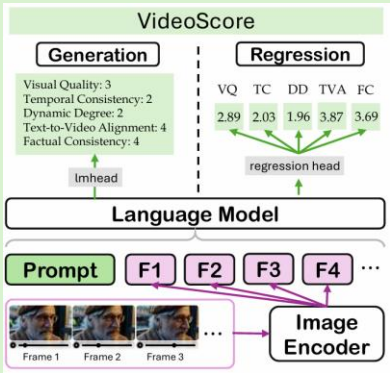
Dataset: VideoFeedback



5 evaluation dimensions
4 levels of quality score
37.6k annotated videos
11 T2V models
(with some real world videos as augmentation)

Pika AnimateDiff
VideoCrafter2

Model: VideoScore



Base Model:
Mantis
Idefics2
VideoLLaVA

Two Versions:
Generation (output natural text of evaluation)
Regression (output floats as quality scores)

Results & Discussion

We test the correlation (Spearman's ρ and Kendall's τ) between VideoScore and human annotation on several benchmarks, compared to various baselines, from MLLM prompting method (query GPT-4o, Gemini-1.5 with the same template) to feature-based metrics (e.g. DINO-sim, CLIP-Score).

Method	Visual Quality	Temporal	Dynamic Degree	Text Alignment	Factual	Avgerage
Random	-3.1	0.5	0.4	1.1	2.9	0.4
Feature-basd automatic metrics						
PIQE	-17.7	-14.5	1.2	-3.4	-16.0	-10.1
BRISQUE	-32.4	-26.4	-4.9	-8.6	-29.1	-20.3
CLIP-sim	21.7	29.1	-34.4	2.0	26.1	8.9
DINO-sim	19.4	29.6	-37.9	2.2	24.0	7.5
SSIM-sim	33.0	30.6	-31.3	4.7	30.2	13.4
MSE-dyn	-20.3	-24.7	38.0	3.3	-23.9	-5.5
SSIM-dyn	-31.4	-29.1	31.5	-5.3	-30.0	-12.9
CLIP-Score	-10.9	-10.0	-14.7	-0.3	-0.3	-7.2
X-CLIP-Score	-3.2	-2.7	-7.3	5.9	-2.0	-1.9
MLLM Propmting						
LLaVA-1.5-7B	9.4	8.0	-2.2	11.4	15.8	8.5
LLaVA-1.6-7B	-8.0	-4.1	-5.7	1.4	0.8	-3.1
Idefics2	4.2	4.5	8.9	10.3	4.6	6.5
Gemini-1.5-Flash	24.1	5.0	20.9	21.3	32.9	20.8
Gemini-1.5-Pro	35.2	-17.2	18.2	26.7	21.6	16.9
GPT-4o	13.6	17.6	28.2	25.7	30.2	23.0
Ours						
VideoScore (gen)	**86.2**	80.3	**77.6**	59.4	82.1	**77.1**
VideoScore (reg)	84.7	**81.5**	68.4	**59.5**	**84.6**	75.7
Δ over Best Baseline	+51.0	+50.9	+39.6	+32.8	+51.7	+54.1

Table 4: Correlation (Spearman's ρ) between model answer and human reference on VideoFeedback-test.

Benchmarks

VideoFeedback-test: 760 videos with human annotation

EvalCrafter Benchmark: Select 3 dimensions that match our evaluation aspects and collect 2500+ videos.

GenAI-Bench and Vbench:
Collect 2100+ videos for GenAI-Bench and select a subset from 5 aspects of Vbench.
We use averaged score of our five dimensions for MLLM prompting baselines and VideoScore to give the preference and calculate the pairwise accuracy as performance indicator.

Case Studies

from VideoFeedback-test 👇

prompt: A robot that throws a stack of paper from a desk

Method	VQ	TC	DD	TVA	FC	Method	VQ	TC	DD	TVA	FC
Human score	3	1	3	3	1						
VideoScore (reg)	2.67	0.81	3.09	2.50	0.80	VideoScore (gen)	3	1	3	3	1
GPT-4o	3	4	2	3	4	Gemini-1.5-Pro	3	1	1	3	3
Gemini-1.5-Flash	3	1	1	3	3	LLaVA-1.6-7B	3	3	3	3	3
LLaVA-1.5-7B	3	3	3	3	2	Idefics1	4	4	3	1	2
PIQE	1	1	1	1	1	DINO-sim	1	1	1	1	1
SSIM-dyn	3	3	3	3	3	CLIP-Score	2	2	2	2	2

from GenAI-Bench 👉