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

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

*Equal Contribution

†Corresponding to: hexuan21@mails.tsinghua.edu.cn, dongfu.jiang@uwaterloo.ca, wenhuchen@uwat

EMNLP 2024

1. Start: evaluate text-to-video (t2v) generation

KLING AI

Prompt: rain on a field of roses

Pika

Prompt: However, what truly set Ezra apart was not his gardening skills but his wisdom and the valuable life lessons he shared with anyone who cared listen

ModelScope

Prompt: Design a motorcycl
drive on the road, with 2
in the front and one w
back

1. Start: evaluate text-to-video (t2v) generation

morph

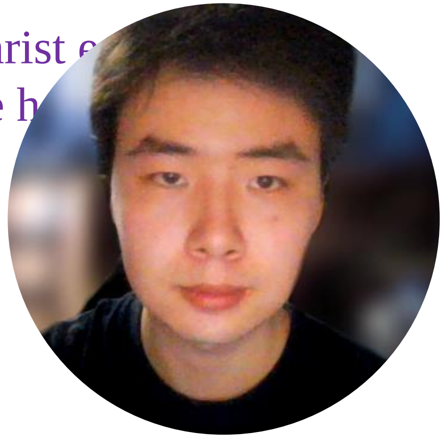
Prompt: One sunny day, as Dorothy cute little girl with long hairs was playing in the fields with Toto, her dog

LVDM

Prompt: A canary bird flucking in cage,pointillism style painting

AnimateDiff

Prompt: Jesus Christ e
an avocado while h



1. Start: evaluate text-to-video (t2v) generation

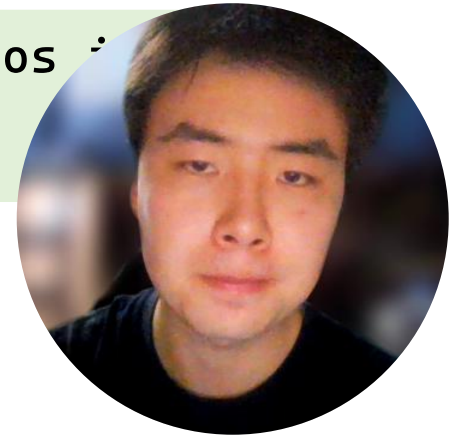
Í 3O 5O3£
G3("≥Ď3

How can we comprehensively evaluate these videos?

provide (1,2,3,4) discrete quality scores for many videos.

finetune Visual LLMs on this large-scale human-annotated data with visual QA format.

Human feedback dataset for AI-generated videos i
required! Then we can train our evaluators!



1. Start: evaluate text-to-video (t2v) generation

Fine-tuning

Dataset: VideoFeedback

First large-scale dataset of **fine-grained human-feedback** dataset for **text-to-video** (T2V) generations

Models: VideoScore

Evaluator model or Reward model for T2V generations

Inference

"An astronaut is riding a horse in the space"

VideoScore

Visual Quality: 2.78
Temporal Consistency: 2
Dynamic Degree: 3
Text-to-Video Alignm
Factual Consistenc

2. VideoFeedback: dataset for t2v evaluation

Evaluation Dimension:

VQ: Visual Quality

TC: Temporal Consistency

DD: Dynamic Degree

TVA: Text-to-Video Alignment

FC: Factual Consistency

Score Scale

4-Real/Perfect

3-Good

2-Avg

1-Bad

VQ	2
TC	2
DD	3
TVA	3
FC	1

Aspect	Definition
Visual Quality (VQ)	the quality of the video in terms of clearness, resolution, brightness, and
Temporal Consistency (TC)	the consistency of objects or humans in video
Dynamic Degree (DD)	the degree of dynamic changes
Text-to-Video Alignment (TVA)	the alignment between the text prompt and the video content
Factual Consistency (FC)	the consistency of the video content with common-sense and factua

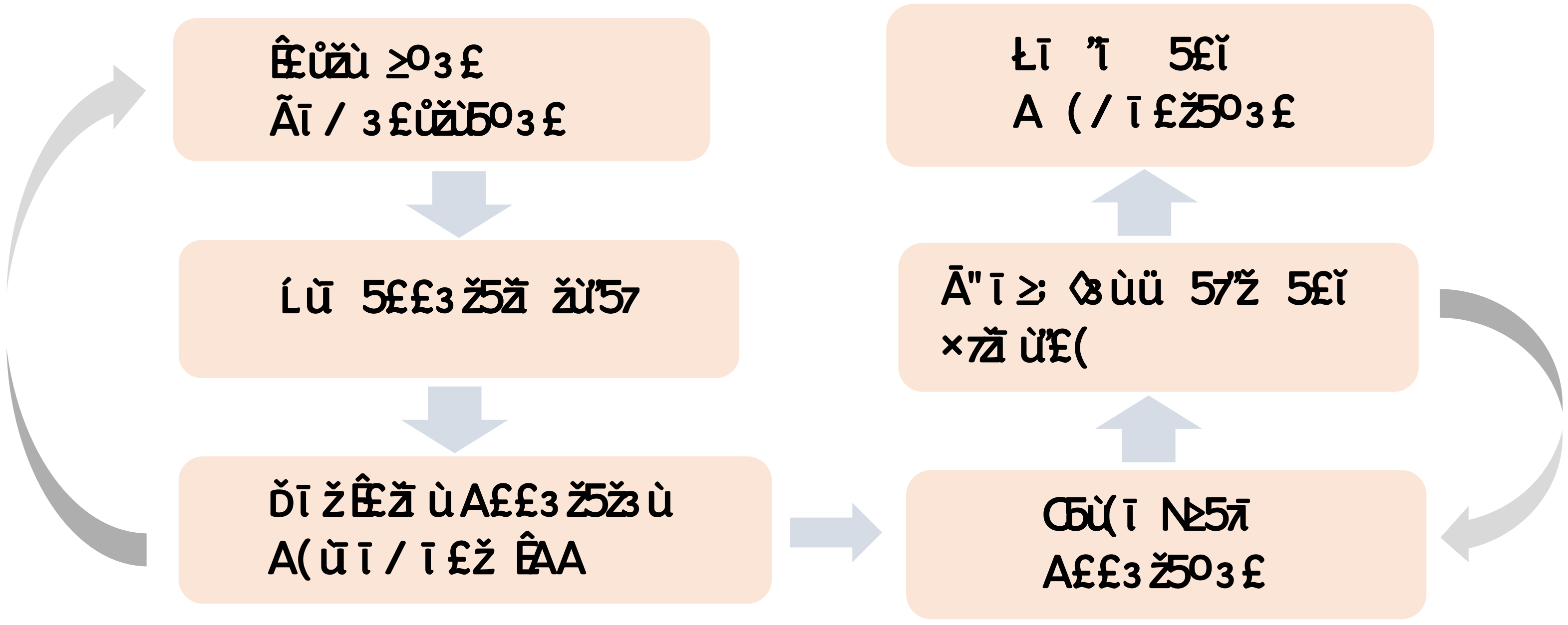
Table 2: The five evaluation aspects of VIDEOFEEDBACK and their definitions.

2. VideoFeedback: dataset for t2v evaluation

Prompts: VidProM (Wang et al, 2024)
Videos: 37.6K videos from 11 T2V models
with various resolution and real videos as augmentation

Base Model or Video Type	Video Source	Total Size	Resolution	Duration	FPS	Score
			Human Annotated Videos			
Pika	VidProM	4.6k	(768, 480)	3.0s	8	[1-4]
Text2Video-Zero (Khachatryan et al., 2023)	VidProM	4.6k	(512,512)	2.0s	8	[1-4]
VideoCrafter2 (Chen et al., 2024a)	VidProM	4.9k	(512, 320)	2.0s	8	[1-4]
ModelScope (Wang et al., 2023a)	VidProM	4.5k	(256, 256)	2.0s	8	[1-4]
LaVie-base (Wang et al., 2023c)	Generated	3.2k	(512, 320)	2.0s	8	[1-4]
AnimateDiff (Guo et al., 2023)	Generated	1.4k	(512, 512)	2.0s	8	[1-4]
LVDM (He et al., 2022)	Generated	3.1k	(256, 256)	2.0s	8	[1-4]
Hotshot-XL (Mullan et al., 2023)	Generated	3.2k	(512, 512)	1.0s	8	[1-4]
ZeroScope-576w (Sterling, 2024)	Generated	2.2k	(256, 256)	2.0s	8	[illegible]
Fast-SVD (Blattmann et al., 2023a)	Generated	1.0k	(1024, 576)	3.0s	8	[illegible]
SoRA-Clip (OpenAI, 2024b)	Collected	0.9k	various	2.0/3.0s	8	[illegible]
			Augmented Videos			
DiDeMo (Hendricks et al., 2017)	Real	2.0k	various	2.0/3.0s	8	[illegible]
Panda70M (Chen et al., 2024b)	Real	2.0k	various	2.0/3.0s	8	[illegible]

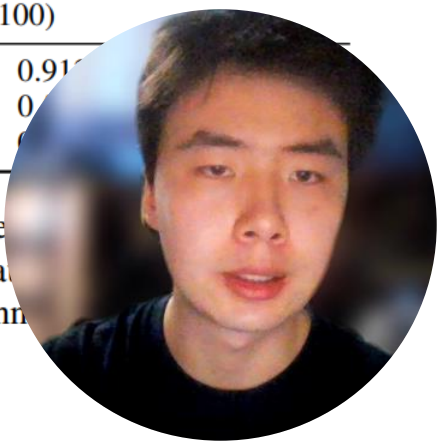
2. VideoFeedback: dataset for t2v evaluation



IAA of Pre-annotation Trial

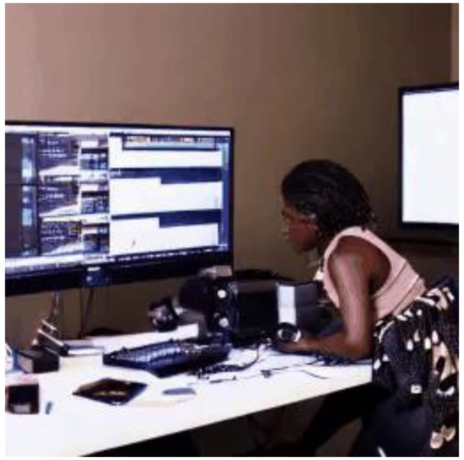
IAA metric	VQ	TC	DD	TVA	FC
Trial 1 (#=30)					
Match Ratio	0.733	0.706	0.722	0.678	0.633
Kappa	0.369	0.414	0.413	0.490	0.265
Alpha	0.481	0.453	0.498	0.540	0.365
Trial 2 (#=100)					
Match Ratio	0.787	0.699	0.91[illegible]	[illegible]	[illegible]
Kappa	0.088	0.562	0.[illegible]	[illegible]	[illegible]
Alpha	0.078	0.579	[illegible]	[illegible]	[illegible]

Table 3: Inter-Annotator Agree
sults considering Matching Ra
pendorff's α on the two trial ann



2. VideoFeedback: dataset for t2v evaluation

VQ	3
TC	3
DD	1
TVA	3
FC	3

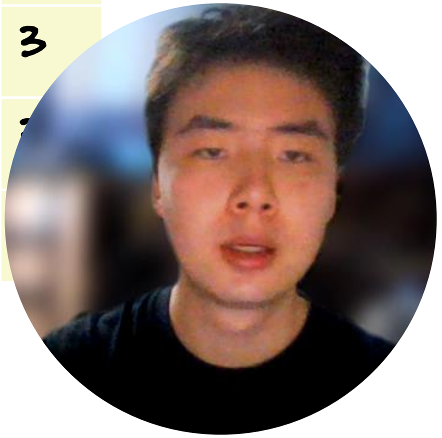


VQ	1
TC	1
DD	3
TVA	3
FC	1

Examples

VQ	3
TC	2
DD	3
TVA	3
FC	1

VQ	1
TC	2
DD	3
TVA	[illegible]
FC	[illegible]



2. VideoFeedback: dataset for t2v evaluation

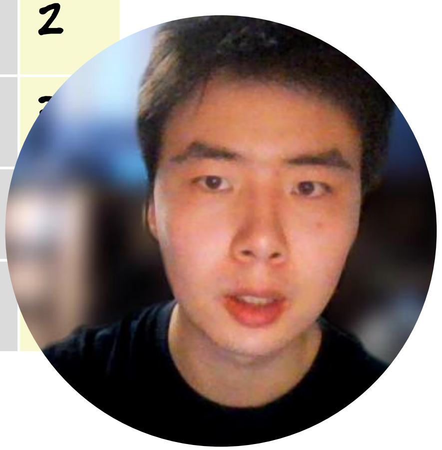
VQ	2
TC	2
DD	3
TVA	3
FC	2

VQ	2
TC	2
DD	3
TVA	2
FC	3

Examples

VQ	1
TC	2
DD	3
TVA	1
FC	2

VQ	2
TC	2
DD	[illegible]
TVA	[illegible]
FC	[illegible]



2. VideoFeedback: dataset for t2v evaluation

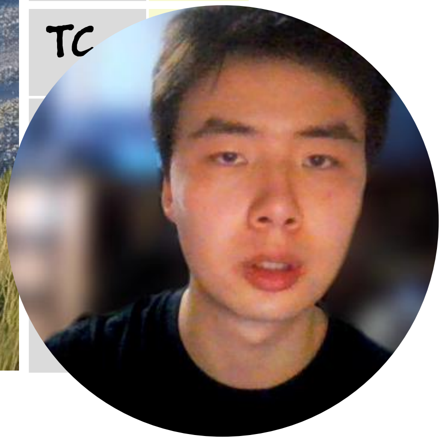
VQ	1
TC	2
DD	3
TVA	2
FC	1

VQ	3
TC	3
DD	3
TVA	3
FC	2

Examples

VQ	4
TC	4
DD	4
TVA	4
FC	4

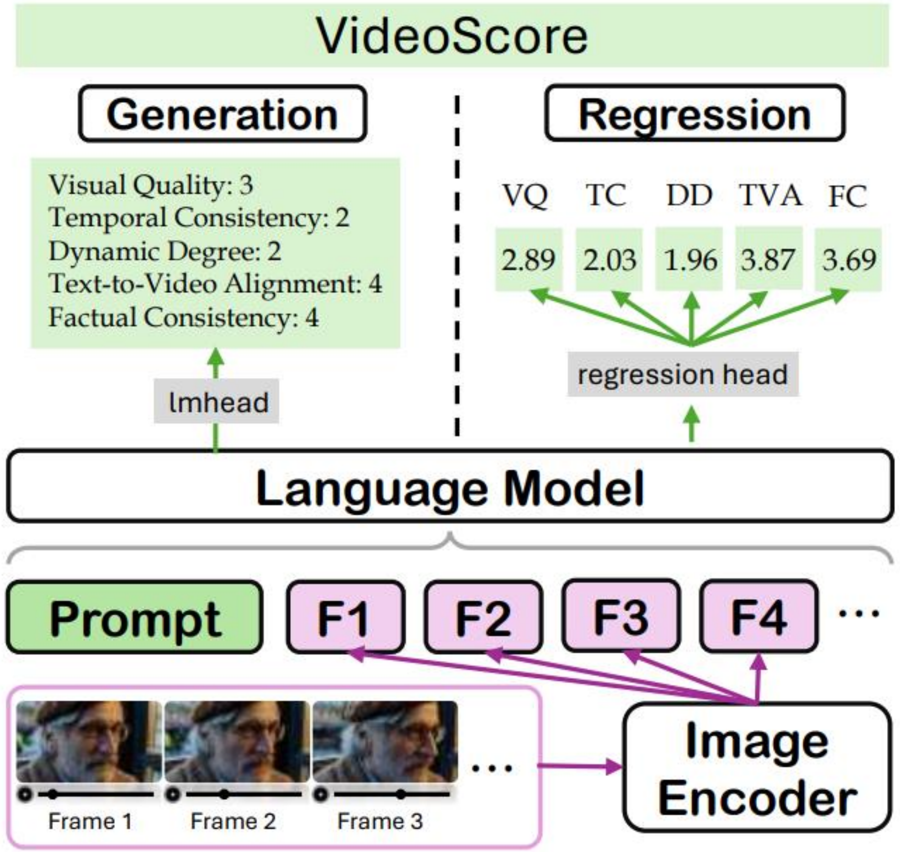
VQ	4
TC	[illegible]



3. VideoScore: fine-grained t2v evaluator

Base Model: Mantis (Main) (Jiang et al,2024)
Ablation model: Idefics2-8B, VideoLLaVA-7B

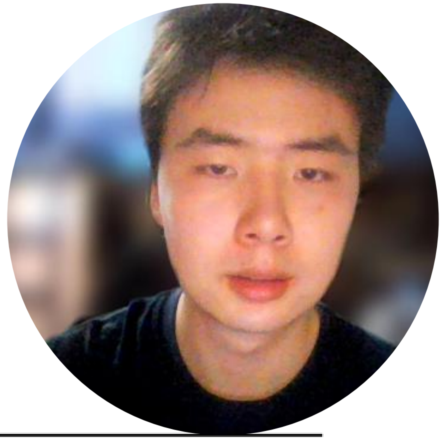
Finetune Data Format: Visual Question Answering
We sample some frames of video, then input them and the text-prompt



Suppose you are an expert in judging and evaluating the quality of AI-generated videos,
please watch the following frames of a given video and see the text prompt for generating the video,
then give scores from 5 different dimensions:
(1) visual quality: the quality of the video in terms of clearness, resolution, brightness, and color
(2) temporal consistency, the consistency of objects or humans in video
(3) dynamic degree, the degree of dynamic changes
(4) text-to-video alignment, the alignment between the text prompt and the video content
(5) factual consistency, the consistency of the video content with the common-sense and factual knowledge

For each dimension, output a number from [1,2,3,4],
in which '1' means 'Bad', '2' means 'Average', '3' means 'Good',
'4' means 'Real' or 'Perfect' (the video is like a real video)
Here is an output example:
visual quality: 4
temporal consistency: 4
dynamic degree: 3
text-to-video alignment: 1
factual consistency: 2

For this video, the text prompt is **"{text_prompt}"**,
all the frames of video are as follows:



3. VideoScore: fine-grained t2v evaluator

Compare with two kinds of baselines on correlation with human-annotated ground truth:

(1) **MLLM Prompting Method**: query MLLMs with the same template and the same sampled frames.

(2) **Feature-based Metric**: e.g. CLIP-sim: average cosine similarity between the CLIP features of adjacent frames.

(See more details in our paper)

VideoScore series | MLLM Prompting Method | Feature-Based Metric

Metric	Final Avg Score ↓	VideoFeedback-test	EvalCrafter	GenAI-Bench	VBench
VideoScore (reg)	**69.6**	75.7	**51.1**	**78.5**	**73.0**
VideoScore-(gen)	55.6	**77.1**	27.6	59.0	58.7
Gemini-1.5-Pro	39.7	22.1	22.9	60.9	52.9
Gemini-1.5-Flash	39.4	20.8	17.3	67.1	52.3
GPT-4o	38.9	23.1	28.7	52.0	51.7
CLIP-sim	31.7	8.9	36.2	34.2	47.4
DINO-sim	30.3	7.5	32.1	38.5	43.3
SSIM-sim	29.5	13.4	26.9	34.1	[illegible]
CLIP-Score	28.6	-7.2	21.7	45.0	[illegible]
LLaVA-1.5-7B	27.1	8.5	10.5	49.9	[illegible]
LLaVA-1.6-7B	23.3	-3.1	13.2	44.5	[illegible]

3. VideoScore: fine-grained t2v evaluator

"A robot that throws a stack of paper from a desk"

VideoScore Inference

Visual Quality:
2.67

Temporal Consistency:
0.81

Dynamic Degree:
3.09

Text-to-Video Alignment:
2.50

Factual Consisten
0.80

3. VideoScore: fine-grained t2v evaluator

"Illustrate a bustling market scene, with fresh produce displayed on stalls, attracting villagers eager to purchase, cartoon style"

VideoScore Inference

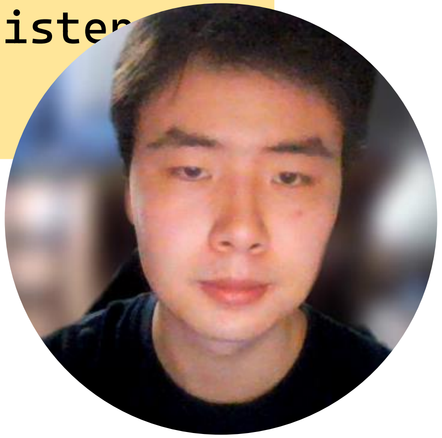
Visual Quality:
1.91
Temporal Consistency:
1.86
Dynamic Degree:
2.84
Text-to-Video Alignment:
2.44
Factual Consisten
1.67

3. VideoScore: fine-grained t2v evaluator

"An astronaut flying in space, oil painting"

VideoScore Inference

Visual Quality:
2.01
Temporal Consistency:
2.63
Dynamic Degree:
2.98
Text-to-Video Alignment:
2.91
Factual Consisten
2.43



3. VideoScore: fine-grained t2v evaluator

"Every day must be Sunday Amusement park inside the school"

VideoScore Inference

Visual Quality:
1.04

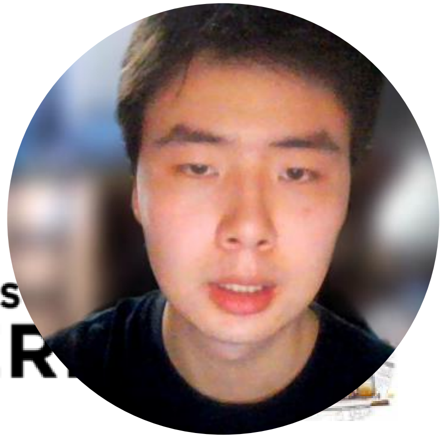
Temporal Consistency:
1.42

Dynamic Degree:
2.95

Text-to-Video Alignment:
1.97

Factual Consisten
1.09

Thanks For Watching



EMNLP 2024