MantisScore: A Reliable Fine-grained Metric for Video Generation

Anonymous ACL submission

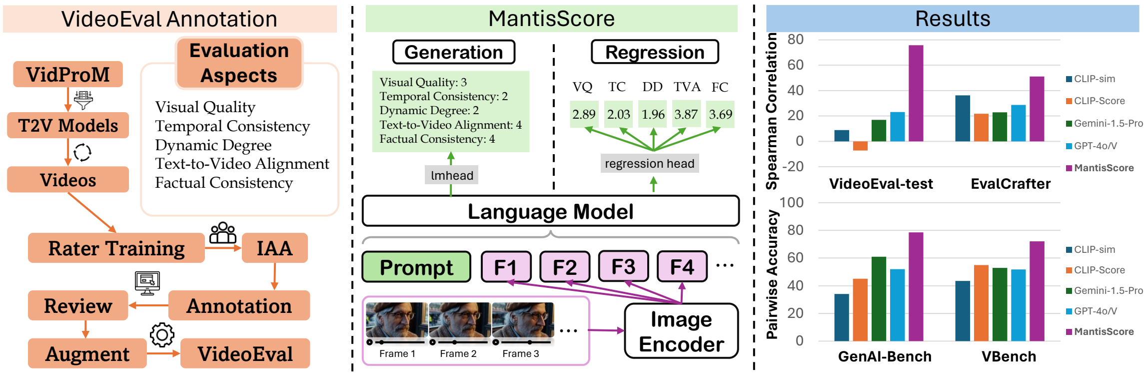

Figure 1: Construction process of VideoEval dataset and illustration of MantisScore.

Abstract

The recent years have witnessed great advances in text-to-video generation. However, the video evaluation metrics have lagged significantly behind, which fails to produce an accurate and holistic measure of the generated videos' quality. The main barrier is the lack of high-quality human rating data. In this paper, we release VideoEval, the first large-scale multi-aspect video evaluation dataset. VideoEval consists of high-quality human-provided ratings for 5 video evaluation aspects on the 37.6K videos generated from 11 existing popular video generative models. We train MantisScore based on VideoEval to enable automatic video quality assessment. Experiments show that the Spearman correlation between MantisScore and humans can reach 77.1 on VideoEval-test, beating the prior best metrics by about 50 points. Further result on the held-out EvalCrafter, GenAI-Bench, and VBench, show that MantisScore is highly generalizable and still beating the prior best metrics by a remarkable margin. We observe that using Mantis as the based model consistently beats that using Idefics2 and VideoLLaVA, and the regression-based model can achieve better results than the generative ones. Due to its high reliability, we believe MantisScore can serve as a valuable tool for accelerate video generation research.

1 Introduction

Powerful text-to-video (T2V) generative models have been exponentially emerging these days. In 2023 and 2024, we have witnessed an array of T2V models like Sora (OpenAI, 2024b), Runway Gen-2 (Esser et al., 2023), Lumiere (Bar-Tal et al., 2024), Pika[1], Luma-AI[2], Kling[3], Emu-video (Girdhar et al., 2023), StableVideoDiffusion (Blattmann et al., 2023a). These models have shown their potential to generate longer-duration, higher-quality, and more natural videos. Despite significant advancements in video generation models, the evaluation metrics of video generation is lagging behind.

The recent literature has adopted a wide range of metrics to do video quality assessments. However, these metrics suffer from the following issues: (1) they can only be used to evaluate visual quality or aesthetics, while failing to capture aspects like motion smoothness, factual consistency, etc. Examples of such metrics include CLIP (Radford et al., 2021b), DINO (Caron et al., 2021), BRISQUE (Mittal et al., 2012a), FVD (Unterthiner et al., 2019), and IS (Salimans et al., 2016). (2) some metrics focus only on a single mean opinion score (MOS), failing to provide fine-grained subscores across different multiple aspects. Examples include T2VQA (Kou et al., 2024b), FastVQA (Wu et al., 2022), and DOVER (Wu et al., 2023). Several works (Ku et al., 2023; Bansal et al., 2024) propose to prompt multi-modal large-language-models (MLLM) like GPT-4o (Achiam et al., 2023) or

[1] https://pika.art/home
[2] https://lumalabs.ai/dream-machine
[3] https://kling.kuaishou.com/

Gemini-1.5 (Reid et al., 2024) to produce multi-aspect quality assessment for given videos. However, our experiments show that they also have low correlation with humans.

The biggest barrier to build reliable video metrics is the lack of high-quality human-annotated dataset. To overcome this barrier, we curate VIDEOEVAL, the first large-scale, multi-aspect video evaluation dataset. We select prompts from VidProM (Wang and Yang, 2024), and use 11 popular text-to-video models, including Pika, Lavie (Wang et al., 2023c), SVD (Blattmann et al., 2023a), etc, to generate videos of various quality based on these prompts. We define five key aspects for evaluation in Table 2, and each aspect is scored from 1 (bad) to 4 (perfect). For annotation, we trained 20 raters to perform a multi-aspect rating over individual generated videos. We have collected ratings for a total of 37.6K videos. We iterate multiple rounds of refinement to ensure a high inter-annotation-agreement (IAA) ratio over 60% for all five aspects.

To build the video evaluator, we select Mantis-Idefics2-8B (Jiang et al., 2024a) as our main backbone model due to its superior ability to handle multi-image and video content, accommodating up to 128 video frames and supporting native resolution. After fine-tuning Mantis on VIDEOEVAL-train, we get our video evaluator, MANTISSCORE. Experiments show that we achieve a Spearman correlation of 77.1 on VIDEOEVAL-test and 59.5 on EvalCrafter (Liu et al., 2023b) for the text-to-video alignment aspect, surpassing the best baseline by 54.1 and 4.4 respectively. The pairwise comparison accuracy gets 78.5 on GenAI-Bench (Jiang et al., 2024b) video preference part, and 72.1 in average on 5 aspects of VBench (Huang et al., 2023), surpassing the previous best baseline by 11.4 and 9.6 respectively. Additional ablation studies with different backbone models confirmed that the Mantis-based metric provides a gain of 12.1 compared to using the Idefics2-based metric. Due to the significant improvement, we believe that MANTISSCORE can serve as the reliable metrics for future video generative models.

2 Related Work

2.1 Text-to-Video Generative Models

Recent progress in diffusion models (Ho et al., 2020; Rombach et al., 2022) has significantly pushed forward the development of Text-to-Video (T2V) generation. Given a text prompt, the T2V generative model can synthesize new video sequences that didn't previously exist (Wang et al., 2023c; OpenAI, 2024b; Chen et al., 2023a, 2024a; Henschel et al., 2024; Bar-Tal et al., 2024). Early diffusion-based video models generally build upon Text-to-Image (T2I) models, adding a temporal module to extend itself into the video domain (Wang et al., 2023c; Chen et al., 2023c). Recent T2V generation models are directly trained on videos from scratch. Among these, models based on Latent Diffusion Models (LDMs) have gained particular attention for their effectiveness and efficiency (Zhou et al., 2022; An et al., 2023; Blattmann et al., 2023b). While the other works used the pixel-based Diffusion Transformers (DiT) also achieve quality results (Gupta et al., 2023; Chen et al., 2023b; OpenAI, 2024b).

2.2 Video Quality Assessment

As the current progress of Text-to-Video generative models leaves it uncertain how close we are to reaching the objective, researchers have worked on evaluation methods to benchmark the generative models. Common methods involve the use of FVD (Unterthiner et al., 2018) and CLIP (Radford et al., 2021a) to evaluate the quality of frames and the text-frames alignment respectively. However, other aspects like subject consistency, temporal consistency, factualness cannot be captured by these metrics. Recent works like VBench (Huang et al., 2023) proposes to use different DINO (Caron et al., 2021), optical flow (Horn and Schunck, 1981) to reflect these aspects. However, the correlation with human judgment is relatively low. For example, most models have subject/background consistency scores over 97% in VBench, which is a massive overestimation of the current T2V models' true capability. Another work EvalCrafter (Liu et al., 2023b) instead resorts to human raters to perform comprehensive evaluation.

A recent work VideoPhy (Bansal et al., 2024) follows VIEScore (Ku et al., 2023) prompt large multi-modal models like Gemini (Reid et al., 2024) and GPT-4o (Achiam et al., 2023) to provide quality assessment. However, our later study shows that these multimodal language models also achieve very low agreement with human raters. A concurrent work T2VQA (Kou et al., 2024a) also proposes to train a quality assessment model on human-annotated video ratings. However, there are a few distinctions. Firstly, our dataset contains ratings for multiple aspects. Secondly, our dataset is 4x larger than the T2VQA dataset. Thirdly, our metric is built on pre-trained video-language foundation models to maximize its performance.

3 VideoEval

This section introduces the construction process of our dataset, VideoEval, for training video evaluators. We start by explaining how we gathered and filtered diverse text prompts for video generation, followed by the video-generation processes using 11 selected text-to-video models. Next, we outline the annotation pipeline that guides raters to score videos across multiple aspects defined in Table 2. We also include supplementary data to enhance robustness. Finally, we summarize the dataset statistics in Table 1, with 760 examples designated as the test set for evaluation.

3.1 Data preparation

Prompt Sources We utilize VidProM (Wang and Yang, 2024), a dataset containing extensive text-to-video pairs from different models. VidProM's video-generation prompts are diverse and semantically rich, derived from real-world user inputs. To create a manageable subset from the 1.04 million unique prompts, we apply two filters: a length filter and an NSFW filter. The length filter eliminates prompts with fewer than 5 words or more than 100 words. The NSFW filter removes prompts with a high probability of containing inappropriate content. After filtering, we perform random down-sampling to obtain a set of 44.5K prompts, 31.6K of them are used in video generation and some videos may have the same text prompt.

Video Generation We select 11 text-to-video (T2V) generative models (shown in Table 1) with various capabilities so that the quality of the generated video ranges from high to low in a balanced way. Some videos are pre-generated in the VidProM dataset, including Pika, Text2Video-Zero (Khachatryan et al., 2023), VideoCrafter2 (Chen et al., 2024a), and ModelScope (Wang et al., 2023a), whereas the others are generated by ourselves or collected from the Internet (i.e. SoRA). To eliminate differences between models in subsequent annotation stage, we normalize the videos into a unified format. First, we standardized the frame rate to 8 fps to address discrepancies in temporal consistency between high and low fps videos. Specifically, for high frame rate model Pika and AnimateDiffusion (Guo et al., 2023) we use frame down sampling, while for low frame rate model like Text2Video-Zero, we employed frame interpolation (Huang et al., 2022) on it. Details are shown in Appendix E. Additionally, we cropped Pika videos to remove the watermark, making them indistinguishable from other models. Ultimately, we obtained 33.6K videos from 11 T2V models, along with their generation prompts.

3.2 Annotation Pipeline

Evaluation Dimensions As discussed in section 1, fine-grained and multi-aspect rating of videos is crucial for enhancing both the reliability and explainability of the video evaluator. Inspired by VBench (Huang et al., 2023) and EvalCrafter (Liu et al., 2023b), and FETV (Liu et al., 2023c), we propose five key dimensions for text-to-video evaluation, detailed in Table 2. These dimensions encompass both low-level vision aspects, such as Visual Quality, which evaluates basic visual impressions, and higher-level aspects, like Text-to-Video Alignment and Factual Consistency, which require a deep understanding of world knowledge, is a capability previous metrics do not have. Besides definition, a checklist for error points for each dimension is also provided to assist the rater in contributing more accurate and consistent rating. Detailed are provided in Table 8.

Annotation We hired 20 expert raters, with each rater performing rating for 1K-2K videos. Our raters are mostly college graduate students. For each aspect, there are three available ratings, 1 (Bad), 2 (Average), and 3 (Good), the score 4 (Perfect) is post-annotated, as described in the subsection 3.3. To ensure the consistency and quality of the annotations, we conducted a system training for each rater. Initially, we conducted a pilot training session with examples of multi-aspect ratings for various videos. Following this, multiple rounds of small-scale annotation were conducted to compute the inter-annotator agreement (IAA) across five aspects, as shown in Table 3. The results indicate a high score-matching ratio for all aspects, along with Fleiss' κ (Fleiss and Cohen, 1973) and Krippendorff's α (Krippendorff, 2011) metrics, with values around 0.4 or 0.5, suggesting sufficient agreement to proceed with large-scale annotation. The annotation process takes roughly 4 weeks to finish.

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	[1-4]
Fast-SVD (Blattmann et al., 2023a)	Generated	1.0k	(1024, 576)	3.0s	8	[1-4]
SoRA-Clip (OpenAI, 2024b)	Collected	0.9k	various	2.0/3.0s	8	[1-4]
Augmented Videos						
DiDeMo (Hendricks et al., 2017)	Real	2.0k	various	2.0/3.0s	8	4
Panda70M (Chen et al., 2024b)	Real	2.0k	various	2.0/3.0s	8	4

Table 1: Statistics of our curated VIDEOEVAL for training video-generation evaluator. It consists of 33.6K human-scored videos across multiple aspects, with 4k real-world videos collected from DiDeMo (Hendricks et al., 2017) and Panda70M (Chen et al., 2024b) as the supplementary data. Ultimately, we get 37.6K high-quality rated videos as the final VIDEOEVAL.

Aspect	Definition
Visual Quality (VQ)	the quality of the video in terms of clearness, resolution, brightness, and color
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 factual knowledge

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

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	0.913	0.570	0.727
Kappa	0.088	0.562	0.565	0.125	-0.089
Alpha	0.078	0.579	0.620	0.205	-0.106

Table 3: Inter-Annotator Agreement (IAA) analysis results considering Matching Ratio, Fleiss' κ, and Krippendorff's α on the two trial annotations.

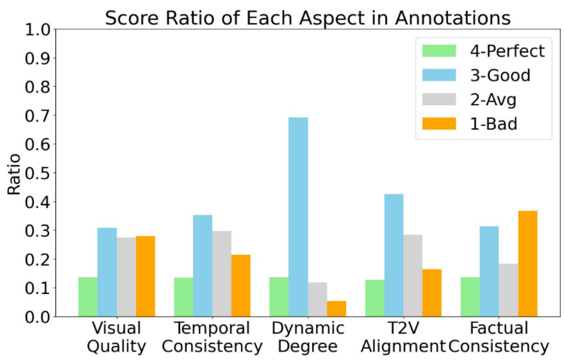

Figure 2: The rating distribution on all the videos.

Review We conduct random checks on human scores during the annotating process. Once we find the exceeded unqualified ratio in certain rater, we promptly communicate with the respective rater and review the annotations for that segment of the video. This helps calibrate the annotation provided by that rater during the relevant period. For example, we found several raters are too lenient and tend to give high scores to unqualified videos. We then step in to make sure they are aligned with our understanding of evaluation dimensions. With periodical random inspection on annotating, we completed the large-scale annotation of 33.6K videos and moved to the data augmentation stage.

3.3 Dataset Augmentation

To enhance the robustness of VIDEOEVAL dataset, we incorporated post-augmentation into the dataset. Firstly, expert raters will review the excellent videos (all aspects are scored 3) again to select perfect ones and raise their scoring to 4 (Perfect) in certain aspects, particularly among the SoRA and FastSVD (Blattmann et al., 2023a) videos.

Additionally, we gather 4k real-world videos from the DiDeMo (Hendricks et al., 2017) and Panda70M (Chen et al., 2024b) with each video accompanied by a text description. We select and cut clips from the ones less than 5 seconds to ensure a strong match between video and its text. We apply similar normalization in subsection 3.1 and also use SSIM and MSE between interval sampled frames to filter out the possible static videos, ensuring the quality in Dynamic Degree. Finally the 4K real videos are scored 4 (perfect) in all aspects.

We plot the rating distributions across each dimension in Figure 2. which is balanced except for Dynamic Degree. We inspected in detail via case study and turned out this distribution is expected. Eventually, we get the final 37.6K examples as the training split of VIDEOEVAL, and reserve 760 examples as VIDEOEVAL-test for evaluation.

4 Experiments

In this section, we describe our experiment setup, including baseline methods for video evaluation, and evaluation benchmarks for video evaluation. We also discuss the training details of MANTISSCORE, and the analysis of our experiment results.

4.1 Baselines

To compare with our evaluator model, we selected two categories of video quality metrics. The first category relies on statistical or neural features for evaluation. These metrics typically assess a single video dimension such as temporal consistency, and then yield a numerical value. The second category employs advanced MLLMs to evaluate videos across multiple dimensions. Extensive literature demonstrates that MLLMs not only excel in generating content on user instructions but also outperform traditional metrics in evaluating AI-generated content (AIGC). All baselines are listed in Table 4.

Feature-Based Metrics

1. Visual Quality. We use two no-reference image quality metrics PIQE (Venkatanath et al., 2015) and BRISQUE (Mittal et al., 2012b). We apply them on all frames of video and take the average score across frames.

2. Temporal Consistency. In this dimension, CLIP-sim (Radford et al., 2021b) and DINO-sim (Caron et al., 2021) are computed as cosine similarities of between adjacent frames features, following VBench (Huang et al., 2023). Additionally, we calculate SSIM between adjacent frames, denoted as SSIM-sim.

3. Dynamic Degree. We uniformly sample four frames from the target video and calculate the average MSE (Mean Square Error) and SSIM (Wang et al., 2004) between adjacent frames in the sample as final score.

4. Text-to-Video Alignment. We include CLIP-Score (Radford et al., 2021b) and X-CLIP-Score (Ma et al., 2022) as metrics in this dimension. CLIP-Score calculates cosine similarity between the feature of each frame and the text prompt and then averages across all frames, while X-CLIP-Score utilizes the feature of video instead of frames.

5. Factual Consistency. It is challenging to find a feature-based metric to determine whether the visual content aligns with common sense. Therefore, we rely on the second category of metrics for this dimension.

We discretized the continuous outputs of these metrics to align with our labeling scores [1, 2, 3, 4]. For instance, for CLIP-sim, values are converted to: '4' if raw output in [0.97, 1], '3' if in [0.9, 0.97), '2' if in [0.8, 0.9) and '1' otherwise. See Table 11 for details.

MLLM Prompting Based Metrics To understand how existing MLLMs perform on the multi-aspect video evaluation task, we designed a prompting template in Table 9 to let them output scores ranging from 1 (Bad) to 4 (Perfect) for each aspect. However, some models, including Idefics2 (Laurençon et al., 2024), Fuyu (Adept AI, 2023), Kosmos-2 (Peng et al., 2023), and CogVLM (Wang et al., 2023b) and OpenFlamingo (Awadalla et al., 2023), fail to give reasonable outputs. We thus exclude them from the tables. MLLMs that follow the output format like LLaVA-1.5 (Liu et al., 2023a), LLaVA-1.6 (Liu et al., 2024), Idefics1 (Laurençon et al., 2023), Google's Gemini 1.5 (Reid et al., 2024), and OpenAI's GPT-4o (OpenAI, 2024a).

4.2 Evaluation Benchmarks

We have included the following benchmarks to evaluate the ability of MANTISSCORE and the above-mentioned baselines on evaluating model generation results.

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	23.1
Ours						
MANTISSCORE (gen)	**86.2**	80.3	**77.6**	59.4	82.1	**77.1**
MANTISSCORE (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 VIDEOEVAL-test.

VIDEOEVAL-test As mentioned in section 3, we split 760 video entries from VIDEOEVAL dataset, which contains 680 annotated videos and 80 augmented videos. We take label prediction accuracy and Spearman's ρ in each dimension as evaluation indicators. For a specific aspect in the VIDEOEVAL-test (e.g. Visual Quality), we use the predicted score from the same aspect to measure the performance for baselines and our models.

GenAI-Bench GenAI-Bench (Jiang et al., 2024b) is a benchmark designed to evaluate MLLM's ability on preference comparison for tasks including text-to-video generation and others. The preference data is taken from GenAI-Arena from user voting. We select the video preference data in our experiments. This involves the MLLM judging which of the two provided videos is generally better, measured by pairwise accuracy. We use the averaged scores of the five aspects for MLLM prompting baselines and our models to give the preference. We compute the correlation between model-assigned preference vs. human preference as our indicator.

VBench VBench (Huang et al., 2023) is a comprehensive multi-aspect benchmark suite for video generative models, where they use a bunch of existing auto-metrics in each aspect. VBench have released a set of human preference annotations on all the aspects, comprising videos by 4 models, including ModelScope (Wang et al., 2023a), CogVideo (Hong et al., 2022), VideoCrafter1 (Chen et al., 2023a), and LaVie (Wang et al., 2023c). We select the subset from 5 aspects of VBench, like technical quality, subject consistency, and so on, to compute the preference comparison accuracy. For each aspect, we subsample 100 unique prompts in the testing. We use the averaged scores of the five aspects for MLLM prompting baselines and our models to predict the preference.

EvalCrafter EvalCrafter (Liu et al., 2023b) is a text-to-video benchmark across four dimensions: Video Quality, Temporal Consistency, Text-to-Video Alignment, and Motion Quality. We focused on the first three ones and gathered 2,541 videos by five models: Pika, Gen2, Floor33 (Floor33, 2024), ModelScope, and ZeroScope (Sterling, 2024). In EvalCrafter, human annotators rated each video on a scale of 1-5, with each scored by three raters. We calculated the average score across raters and normalized it to [0, 1]. After inference on benchmark videos, we excluded "Dynamic Degree" and "Factual Consistency" to match EvalCrafter's dimensions. Finally, we used Spearman's ρ in each dimension as an indicator.

Benchmark →	GenAI-Bench	VBench				
Model ↓ Sub-Aspect →	Video Preference	Technical Quality	Subject Consistency	Dyanmic Degree	Motion Smoothness	Overall Consistency
Random	37.7	44.5	42.0	37.3	40.5	44.8
Feature-based Automatic Metrics						
PIQE	34.5	60.8	44.3	71.0	45.3	53.8
BRISQUE	38.5	56.7	41.2	75.5	41.2	54.2
CLIP-sim	34.1	47.8	46.0	34.8	44.7	44.2
DINO-sim	31.4	49.5	51.2	24.7	55.5	41.7
SSIM-sim	28.4	30.7	46.2	24.5	54.2	27.2
MSE-dyn	34.2	32.8	31.7	81.7	31.2	39.2
SSIM-dyn	38.5	37.5	36.3	**84.2**	34.7	44.5
CLIP-Score	45.0	57.8	46.3	71.3	47.0	52.2
X-CLIP-Score	41.4	44.0	38.0	51.0	28.7	39.0
MLLM Prompting						
LLaVA-1.5-7B	49.9	42.7	42.3	63.8	41,33	8.8
LLaVA-1.6-7B	44.5	38.7	26.8	56.5	28.5	43.2
Idefics1	34.6	20.7	22.7	54.0	27.3	33.7
Gemini-1.5-Flash	67.1	52.3	49.2	64.5	45.5	49.9
Gemini-1.5-Pro	60.9	56.7	43.3	65.2	43.0	56.3
GPT-4o	52.0	59.3	49.3	46.8	42.0	60.8
Ours						
MantisScore (gen)	59.0	64.2	57.7	55.5	54.3	61.5
MantisScore (reg)	**78.5**	**78.2**	**71.5**	68.0	**74.0**	**69.0**
Δ over Best Baseline	+11.4	+17.3	+20.3	-16.2	+18.5	+8.2

Table 5: Pairwise preference accuracy on GenAI-Bench (Jiang et al., 2024b) and VBench (Huang et al., 2023). For MLLM prompting and our method, we averaged the five aspect scores defined in Table 2 as the score for each video in the comparison, where the higher one deemed the winner. The table below shows the accuracy of each method by comparing these computed scores with human annotations of "Win," "Tie," and "Lost" for the two videos.

Method	Visual	Temporal	Text Align
Random	-2.0	1.4	-0.9
EvalCraft (GPT-4V)	**55.4**	**56.7**	32.3
Feature-based Automatic Metrics			
PIQE	0.5	-3.3	-0.9
BRISQUE	6.4	-1.3	6.7
CLIP-sim	36.0	53.5	19.2
DINO-sim	30.6	50.3	15.3
SSIM-im	32.4	36.9	11.4
MSE-dyn	-15.4	-27.5	-8.1
SSIM-dyn	-32.6	-33.9	-12.6
CLIP-Score	18.7	11.5	35.0
X-CLIP-Score	12.2	3.1	24.5
MLLM Prompting			
LLaVA-1.5-7B	13.4	15.6	2.6
LLaVA-1.6-7B	12.2	8.5	18.9
Idefics1	1.5	-1.5	0.8
Gemini-1.5-Flash	34.9	-27.8	44.8
Gemini-1.5-Pro	37.8	-24.1	55.1
GPT-4o	32.9	12.5	40.7
Ours			
MantisScore (gen)	20.8	51.3	10.7
MantisScore (reg)	42.4	51.3	**59.5**
Δ over Best Baseline	-13.1	-5.4	4.4

Table 6: Spearman's Correlation (ρ) of MantisScore on EvalCrafter (Liu et al., 2023b)

4.3 Training Details

For MantisScore, We use two scoring methods: generative scoring and regression scoring. Generative scoring involves training the model to output fixed text forms, from which aspect scores are extracted using regular expressions. These scores are integers corresponding to human annotation scores. In contrast, regression scoring replaces the language model head with a linear layer that outputs 5 logits representing scores for each aspect. Regression scoring is trained using MSE loss.

We select Mantis-Idefics2-8B (Jiang et al., 2024a) as the base model, which can accommodate 128 video frames at most. The learning rate is set to 1e-5. Each model is trained for 1 epoch on 8 A100 (80G) GPUs, finishing in 6 hours.

4.4 Results

We report the Spearman correlation results on the VideoEval-test and EvalCrafter in Table 4 and Table 6, respectively. For the preference comparison on videos, we report the pairwise accuracy on the GenAI-Bench and VBench in Table 5.

Base Model	Scoring Type	VIDEOEVAL *	EvalCrafter*	GenAI-Bench	VBench*	Average
VideoLLaVA-7B	Generation	71.9	9.8	42.6	46.5	42.7
Idefics2-8B	Generation	73.9	11.3	50.7	53.9	47.5
Mantis-Idefics2-8B	Generation	77.1	27.6	59.0	58.7	55.6
Idefics2-8B	Regression	73.9	17.4	74.5	64.4	57.5
Mantis-Idefics2-8B	Regression	**75.7**	**51.1**	**78.5**	**73.0**	**69.6**

Table 7: Ablation study on the base model and scoring function for MANTISSCORE. "*" means that we take the average of Spearman correlation or pairwise accuracy across the multiple aspects of the benchmark. The highest numbers are bold for each benchmark, and the second are underlined.

MANTISSCORE achieves the SoTA performance On the VIDEOEVAL-test, MANTISSCORE gets an average of 54.1 improvements on all the five aspects compared to the baseline GPT-4o. What's more, on the EvalCrafter benchmark, MANTISSCORE (reg) has 4.4 improvements on text-to-video alignment. For pairwise preference comparison, MANTISSCORE also gets 78.5 accuracy on GenAI-Bench, surpassing the second-best Gemini-1.5-Flash by 11.4 points. on the Vbench, our model archives the highest pairwise accuracy on 4 out of 5 aspects from VBench, with an average of 16.1 improvements.

Feature-based Automatic Metrics are limited While some feature-based automatic metrics are good at a single aspect, they might fail to evaluate well on others. For example, on the VIDEOEVAL-test, the correlation scores of SSIM-dyn and MSE-dyn achieve 31.5 and 38.0 for the dynamic degree aspect, but they both get a negative correlation for others. Besides, PIQE, BRISQUE, CLIP-Score, and X-CLIP-Score get nearly all negative correlations for all 5 aspects. This proves the image quality assessment metrics cannot be easily adapted to the video quality assessment task.

4.5 Ablation Study

We conducted an ablation study on the base model selection and scoring types by training different variants on VIDEOEVAL. Results shown in Table 7.

Base model ablation To investigate the effects of changing the base model, we have trained different variants with VideoLLaVA-7B and Idefics2-8B as the base models. Since VIDEOEVAL-test, EvalCrafter, and VBench both have multiple aspects in the benchmarks, we take the average score across these aspects and report the general performance in Table 7. The results show that the Video-LLaVA-based version gets the worst performance on the four benchmarks, even if it is specifically designed for video understanding. The Idefics2-8B-based version has marginal improvements compared to the VideoLLaVA. After changing to Mantis-Idefics2-8B, the scores on the four benchmarks keep improving from 47.5 to 55.6 on average. When the scoring type is regression, the mantis-based version is still better than the Idefics2-based version by 12.1 points. Therefore, we select the Mantis-based version as the final choice.

Regression scoring or generative scoring? The primary difference between regression scoring and generative scoring is that regression scoring can give more fine-grained scores instead of just the four labels. Results on EvalCrafter, GenAI-Bench, and VBench all indicate that using regression scoring can consistently improve the Spearman correlation or the pairwise comparison accuracy. For example, on GenAI-Bench, MANTISSCORE (reg) achieves 78.5 accuracy, which is higher than the 59.0 of the MANTISSCORE (gen). The results are similar for the other benchmarks. We thus conclude that regression scoring with more fine-grained scores is a better choice.

5 Conclusion

In this paper, we introduce MANTISSCORE, which is trained on our meticulously curated dataset VIDEOEVAL for video evaluation. We hired 20 expert raters to annotate the 37.6K videos generated from 11 popular text-to-video generative models across 5 key aspects, Visual Quality, Temporal Consistency, Dynamic Degree, Text-to-Video Alignment and Factual Consistency. Our IAA match ratio gets more than 60%. We test the performance of MANTISSCORE using Spearman correlation on VIDEOEVAL-test and EvalCrafter, and using pairwise comparison accuracy on GenAI-Bench and VBench. The results show that MANTISSCORE consistently gets the best performance, surpassing the powerful baseline GPT-4o and Gemini 1.5 Flash/Pro by a large margin. Our work highlights the importance of using MLLM for video evaluation due to its rich world knowledge and the high-quality rating dataset across multiple aspects.

References

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

Adept AI. 2023. Fuyu-8B: A Multimodal Architecture for AI Agents. https://www.adept.ai/blog/fuyu-8b.

Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. 2023. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2024. Videophy: Evaluating physical commonsense for video generation.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. 2023a. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*.

Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. 2023b. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv preprint arXiv:2312.04557*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024b. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023c. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Floor33. 2024. Floor33 pictures: Ai video generator.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. 2023. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. 2023. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. *Preprint*, arXiv:1708.01641.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2024. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Berthold KP Horn and Brian G Schunck. 1981. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. 2022. Real-time intermediate flow estimation for video frame interpolation. *Preprint*, arXiv:2011.06294.

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. 2023. Vbench: Comprehensive benchmark suite for video generative models. *ArXiv*, abs/2311.17982.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024a. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. 2024b. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*.

Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. 2024a. Subjective-aligned dataset and metric for text-to-video quality assessment. *ArXiv*, abs/2403.11956.

Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. 2024b. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *Preprint*, arXiv:2312.14867.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *ArXiv*, abs/2405.02246.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2023b. Evalcrafter: Benchmarking and evaluating large video generation models.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2023c. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Preprint*, arXiv:2311.01813.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012a. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012b. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.

John Mullan, Duncan Crawbuck, and Aakash Sastry. 2023. Hotshot-XL.

OpenAI. 2024a. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-06-15.

OpenAI. 2024b. Video generation models as world simulators.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Spencer Sterling. 2024. Zeroscope v2.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation.

N. Venkatanath, D. Praneeth, Maruthi Chandrasekhar Bh, Sumohana Channappayya, and Swarup Medasani. 2015. Blind image quality evaluation using perception based features. *2015 21st National Conference on Communications, NCC 2015*.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023a. Modelscope text-to-video technical report. *ArXiv*, abs/2308.06571.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023b. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079.

Wenhao Wang and Yi Yang. 2024. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *ArXiv*, abs/2403.06098.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2023c. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, pages 538–554.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.

A Ethical Statement

This work fully complies with the ACL Ethics Policy. We declare that there are no ethical issues in this paper, to the best of our knowledge.

B Risks and Limitation

Although we have designed systematic pipelines to recruit expert raters and annotate the video evaluation scores, we still find out that some annotations contain errors and may harm the overall quality of the dataset. Our IAA score computation is only based on a small number of trial examples and, thus might not represent the actual IAA of the whole annotations.

Besides, while MANTISSCORE is proven to be able to effectively give reasonable scores on our defined five aspects, it can still sometimes output wrong scores that do not match our expectations. We admit this drawback and list that as one of our future works.

C Dataset Licence

We have used VidProM (Wang and Yang, 2024) to collect the prompts used for video generation, whose usage LICENSE is CC BY-NC 4.0 license. For other evaluation datasets, We did not find license for EvalCrafter (Liu et al., 2023b) human annotations. GenAI-Bench (Jiang et al., 2024b) is under MIT licence, and VBench (Huang et al., 2023) is under Apache 2.0 license. We are thus able to utilize these datasets in our experiments.

We also release our curated dataset, VIDEOEVAL, under MIT license to contribute to the video evaluation dataset.

D Annotator Management

During the annotation, we have recruited 20 expert raters, where 14 of them are undergraduate or graduate students, who will become one of the authors of our paper, and the rest of them are assured to be paid with decent salary.

E Video Format Normalizing Details

To mitigate difference of videos format from different generative models, we normalize the frame rate of all the generated videos to 8 fps (frames per second). Specifically, for high frame rate model Pika and AnimateDiffusion (Guo et al., 2023), we use uniform down-sampling to normalize Pika from 24 fps to 8fps, and AnimateDiffusion from 23 fps to 8 fps. For low frame rate model Text2Video-Zero (Khachatryan et al., 2023), we use video frame interpolation model RIFE (Huang et al., 2022) to interpolate frames, adding the frame rate from 4 fps to 8 fps. For real-world videos from DiDeMo (Hendricks et al., 2017) and Panda70M (Chen et al., 2024b) in post augmentation of VIDEOEVAL, we use the same down-sampling as Pika and AnimateDiffusion to reduce their frame rate from 30 fps to 8 fps.

Additionally, since video from Pika are always attached a watermark "PIKA-LABS", we cropped all the Pika videos from the resolution of (1088, 640) to (768, 480), making Pika video indistinguishable from videos from other models.

F Annotation Details

Additional annotation details are put in this section for the reference.

Firstly we show the user interface of our annotating website in Figure 3 and Figure 4. In both welcome page and working page, we list the definition and a checklist of error points in five evaluation dimensions, as shown in Table 8. Additionally we also provide many Good/Average/Poor videos as examples in each dimension for raters to quickly understand each dimension and align well with our understanding.

G Prompting Template

In process of training Mantis (Jiang et al., 2024a) for generation scoring and the testing with "MLLM Prompting" baselines, we use the same prompt template provided in Table 9.

For training Mantis with regression scoring, we make modification to the above template accordingly, instructing model to output a float number ranges from 1.0 to 4.0, as shown in Table 10.

H Feature-based Baselines Discretization

As described in subsection 4.1, we employ several statistical or neural feature-based metrics as baselines for comparison with our model. The continuous float-format outputs of these metrics are discretized into labels [1, 2, 3, 4], aligning with our annotation data format. The discretization rules are presented in Table 11. Metrics with a ↑ symbol indicate that higher values are better, while those with a ↓ symbol indicate that lower values are better.

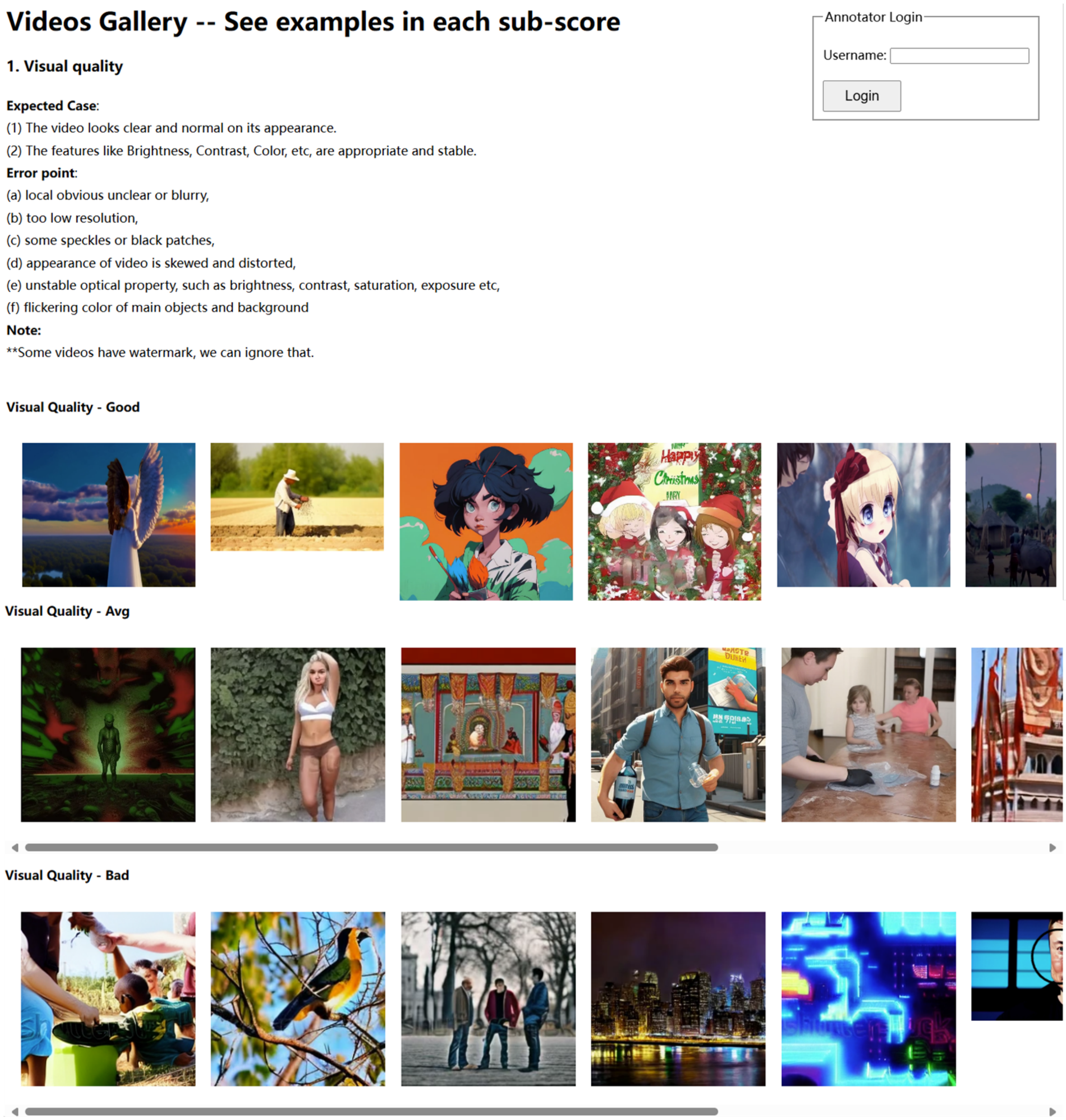

Figure 3: Welcome Page of our video annotating website, with definition, checklist for error points and diverse video examples.

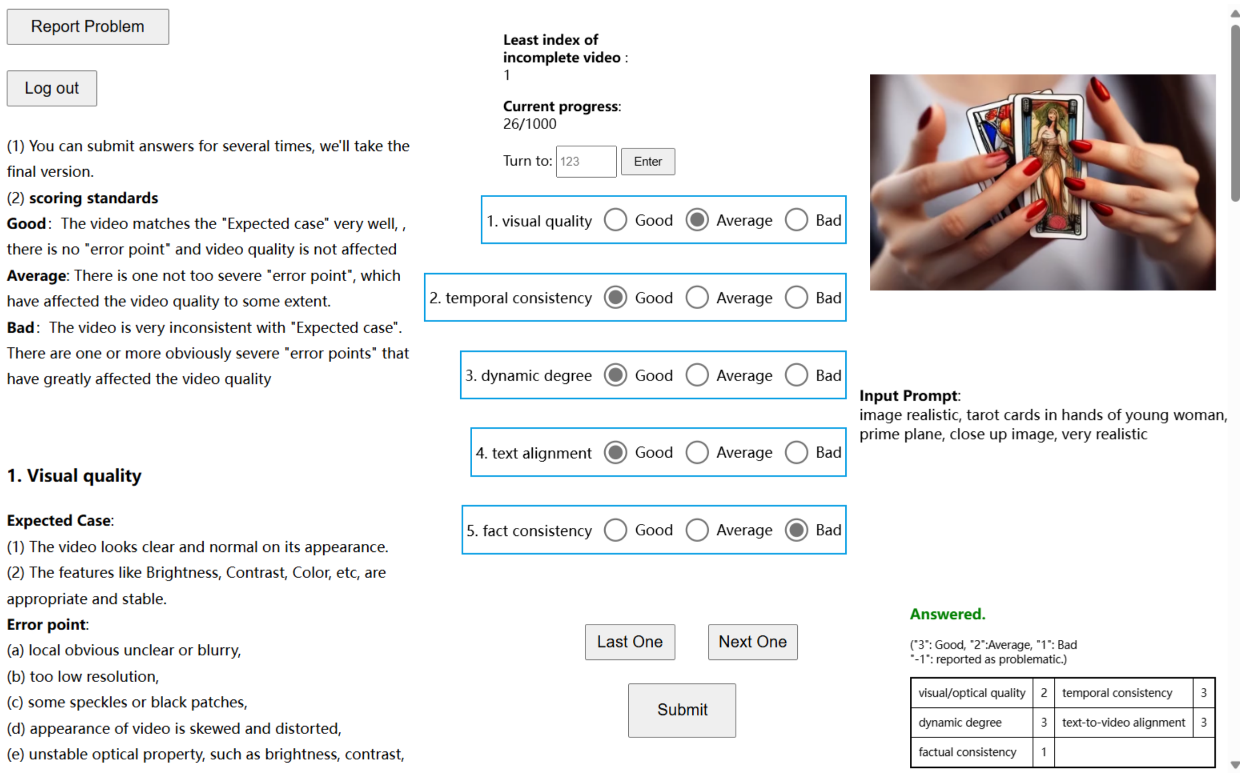

Figure 4: Working page of our video annotating website

I Case study of VIDEOEVAL

We showcase the annotations examples in Figure 5. The first example depicts a clear video of a woman with her hair moving, thus scoring 3 in all 5 aspects. The second example shows a distorted video, thus scoring 1 across all the aspects except the dynamic degree. We further analyzed the correlations between the designed aspects in Figure 6. We found that visual quality achieves a high correlation of 0.6 with temporal consistency, while dynamic degree has a very low correlation with all other aspects.

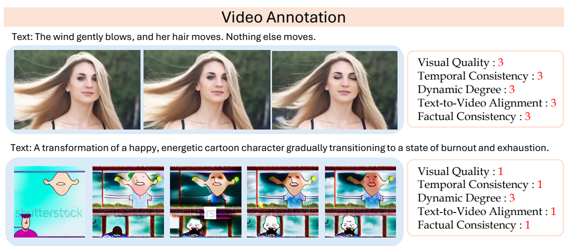

Figure 5: Example of annotations. Each video has a text description and is rated for the 5 aspects.

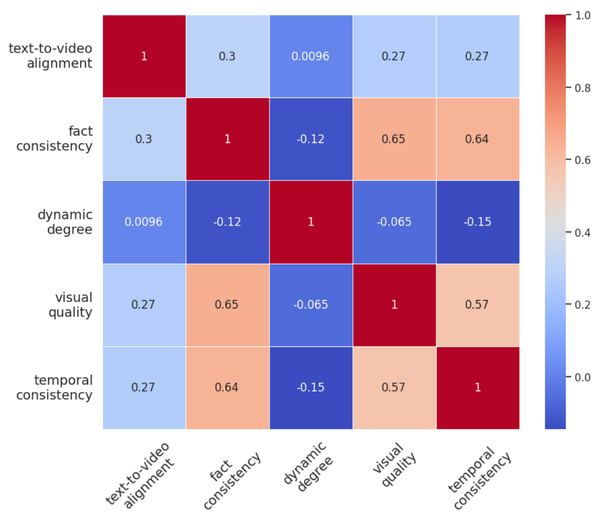

Figure 6: Correlation study on the evaluation aspects.

Evaluation Aspect	Detailed Description for Annotation
Visual Quality	Expected Case:(1) The video looks clear and normal on its appearance. (2) The features like Brightness, Contrast, Color, etc, are appropriate and stable. Error point:(a) local obvious unclear or blurry, (b) too low resolution, (c) some speckles or black patches, (d) appearance of video is skewed and distorted, (e) unstable optical property, such as brightness, contrast, saturation, exposure etc, (f) flickering color of main objects and background Note:Some videos have watermark, we can ignore that.
Temporal Consistency	Expected Case: (1) The main objects, main characters and overall appearance are consistent across the video. (2) The appearance of video as well as the movements of humans and objects are smooth and natural. Error points: (a) The person or object suddenly disappears or appears (b) The type or class of objects has obvious changes (c) There is an obvious switch in the screen shot (d) the appearance of video or movements in it is laggy and un-smooth, (e) local deformation or dislocation of human or objects due to the motion (for large scale deformation, the video should also be rated as bad in "1. visual quality"), Note: For a video almost static or with small dynamic degree, as long as it does not have error points, then it should be scored as good.
Dynamic Degree	Expected Case: (1) The video is obviously not static, the people or objects or the video screen is dynamic. (2) The video can be easily distinguished from a static image. Note: You are supposed to focus on only dynamic degree, regardless of the visual quality and video content
Text-to-Video Alignment	Expected Case: The characters, objects, motions, events etc. that are mentioned in text input prompts all exist reasonably. Error points: (a) The people and objects in prompt do not appear in video (b) The actions and events in prompt do not appear in video (c) The number, size, shape, color, state, movement and other attributes of the objects in the video do not match the prompt (d) Text mentioned in prompt is not displayed correctly in the video, such as "a placard saying 'No Smoking'" but "No Smoking" is not spelled correctly in the video (e) The video format (such as width, height, screen ratio, duration) does not match the format in prompt.
Factual Consistency	Expected Case: (1) Overall appreance and motion are consistent with our common-sense, physical principles, moral standards, etc. Error points: (a) static ones: Content in video goes against common sense in life, such as lighting a torch in the water, standing in the rain but not getting wet, etc. (b) static ones: The size, color, shape and other basic properties of objects violate scientific principles (c) dynamic ones: The overall movement of people or objects violates common-sense and laws of physics, such as spontaneous upward movement against gravity, abnormal water flow, etc. (d) dynamic ones: Partial movements of people or objects violate common-sense and laws of physics, such as the movement of hands or legs is anti-joint, etc. Notes: Relation with '5. text-to-video alignment': Some text prompts express fictional and unrealistic content, for example, "a dog plays the guitar in the sky" or "an astronaut rides a horse in space". In this case, regardless of the veracity of the text prompt, you should only consider whether the other content in the video makes sense.

Table 8: Expected cases and error cases for each aspect that annotators can see during the annotation.

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:

Table 9: Prompting template in generation format used for MANTISSCORE training and the MLLM prompting baselines

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 float number from 1.0 to 4.0,
higher the number is, better the video performs in that dimension,
the lowest 1.0 means Bad, the highest 4.0 means Perfect/Real (the video is like a real video)
Here is an output example:
visual quality: 2.24
temporal consistency: 3.89
dynamic degree: 3.17
text-to-video alignment: 1.86
factual consistency: 2.16

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

Table 10: Prompting template used for the MLLM prompting baseline and MANTISSCORE training

Dimension	Metric	1 (Bad)	2 (Avg)	3 (Good)	4 (Perfect)
Visual Quality	PIQE$\downarrow$	$[50, \infty)$	[30,50)	[15,30)	[0,15]
	BRISQUE$\downarrow$	$[50,\infty)$	[30,50)	[10,30)	[0,10]
Temporal Consistency	CLIP-sim$\uparrow$	[0,0.80)	[0.80,0.90)	[0.90,0.97)	[0.97,1]
	DINO-sim$\uparrow$	[0,0.75)	[0.75,0.85)	[0.85,0.95)	[0.95,1]
	SSIM-sim$\uparrow$	[0,0.6)	[0.6,0.75)	[0.75,0.9)	[0.9,1]
Dynamic Degree	MSE-dyn$\uparrow$	[0,100]	[100,1000)	[1000,3000)	$[3000,\infty)$
	SSIM-dyn$\downarrow$	[0.9,1]	[0.7,0.9)	[0.5,0.7)	[0,0.5)
Text-to-Video Alignment	CLIP-Score$\uparrow$	[0.2,0.27)	[0.27,0.31)	[0.31,0.35)	[0.35,0.4]
	X-CLIP-Score$\uparrow$	[0,0.15)	[0.15,0.23)	[0.23,0.30)	[0.30,1]

Table 11: Discretization rules for featured-based baselines.