

Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment

Tengchuan Kou
Shanghai Jiao Tong University
Shanghai, China
2213889087@sjtu.edu.cn

Xiaohong Liu*
Shanghai Jiao Tong University
Shanghai, China
xiaohongliu@sjtu.edu.cn

Zicheng Zhang
Shanghai Jiao Tong University
Shanghai, China
zzc1998@sjtu.edu.cn

Chunyi Li
Shanghai Jiao Tong University
Shanghai, China
lcysyzxdc@sjtu.edu.cn

Haoning Wu
Nanyang Technological University
Singapore, Singapore
haoning001@e.ntu.edu.sg

Xiongkuo Min
Shanghai Jiao Tong University
Shanghai, China
minxiongkuo@sjtu.edu.cn

Guangtao Zhai
Shanghai Jiao Tong University
Shanghai, China
zhaiguangtao@sjtu.edu.cn

Ning Liu*
Shanghai Jiao Tong University
Shanghai, China
ningliu@sjtu.edu.cn

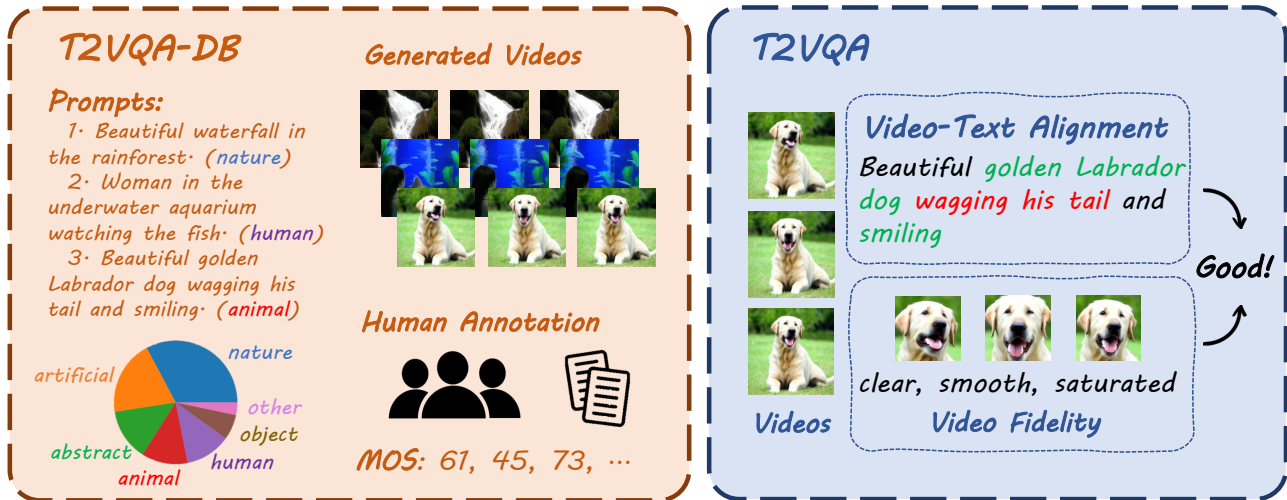


Figure 1: Overview of the proposed T2VQA-DB and T2VQA. T2VQA-DB has the largest scale among existing T2V datasets. T2VQA achieves the SOTA performance in evaluating the quality of text-generated videos.

Abstract

With the rapid development of generative models, AI-Generated Content (AIGC) has exponentially increased in daily lives. Among them, Text-to-Video (T2V) generation has received widespread

attention. Though many T2V models have been released for generating high perceptual quality videos, there is still lack of a method to evaluate the quality of these videos quantitatively. To solve this issue, we establish the largest-scale Text-to-Video Quality Assessment DataBase (**T2VQA-DB**) to date. The dataset is composed of 10,000 videos generated by 9 different T2V models, along with each video’s corresponding mean opinion score. Based on T2VQA-DB, we propose a novel transformer-based model for subjective-aligned Text-to-Video Quality Assessment (**T2VQA**). The model extracts features from text-video alignment and video fidelity perspectives, then it leverages the ability of a large language model to give the prediction score. Experimental results show that T2VQA outperforms existing T2V metrics and SOTA video quality assessment models. Quantitative analysis indicates that T2VQA is capable of giving subjective-align predictions, validating its effectiveness. The dataset and code are available at <https://github.com/QMME/T2VQA>.

*Corresponding authors.

CCS Concepts

• **Computing methodologies** → **Modeling methodologies**.

Keywords

Text-to-video dataset, Video quality assessment, Text-to-video generation

ACM Reference Format:

Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. 2024. Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28–November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3664647.3680868>

1 Introduction

Video generation, or video synthesis, has been fully developed in the past few years. Text-to-Video (T2V) generation is one of the most studied fields, where a user provides a text description as the guidance for video generation. With the thriving of diffusion-based models, high-fidelity videos can be generated. However, the quality of text-generated videos is diverse which affects the experience quality of subjects. Therefore, a subjective-aligned quality assessment method for them is needed. Unfortunately, existing Video Quality Assessment (VQA) models are unable to accomplish the task well. On the one hand, distortions brought by T2V generation models, such as jitter effect, irrational objects, are different from distortions in natural videos. On the other hand, traditional VQA models do not take text-video alignment into consideration, which is a significant evaluation perspective for text-generated videos.

Besides, the most used metrics for T2V generation, such as IS [38], FVD [44], and CLIPSim [48], fail to reflect real user preferences. IS uses the Inception Network [42] to generate a distribution that reflects image/video quality and diversity. It has been criticized for its inability to evaluate image/video quality precisely. FVD compares the I3D feature [4] distributions of the generated and natural video pair. The drawback of FVD is that obtaining the reference natural video is usually impractical. CLIPSim takes advantage of CLIP [34] to calculate the similarity between the original text and the generated video content. However, it only considers text-video alignment from the image level, excluding the temporal information and perceptual video quality.

To facilitate the development of a more comprehensive and accurate metric, we establish the largest-scale subjective T2V dataset to date, named Text-to-Video Quality Assessment DataBase (**T2VQA-DB**). The dataset contains 10,000 videos generated by 9 representative T2V models using 1,000 text prompts. We also collect each video’s Mean Opinion Score (MOS) by conducting a subjective experiment, where 27 subjects score the overall quality of the generated videos. Fig. 2 shows video frames generated by the 9 T2V models from the prompt “Sunset on the sea”. And we give more detailed overviews of three examples. We anticipate that the T2VQA-DB will benefit the development of subsequent models.

Based on T2VQA-DB, we propose a novel model equipped with multi-modality foundation models for better Text-to-Video Quality Assessment (**T2VQA**). The model utilizes BLIP [29] and Video Swin Transformer (Swin-T) [32] to extract features from text-video

Table 1: Summary of T2V datasets. [Bold: the best].

Name	Videos	Prompts	Models	Annotators
Chivileva’s [5]	1,005	201	5	24
ECTV [30]	5,600	700	8	7
VBench [17]	6,984	1,746	4	-
FETV [31]	2,476	619	4	3
TVGE [57]	2,543	512	5	10
T2VQA-DB (ours)	10,000	1,000	9	27

alignment and video fidelity perspectives respectively. The features are fused through a cross-attention module, then they are fed into a frozen Large Language Model (LLM) to regress the predicted score. We train and test T2VQA as well as other VQA models on T2VQA-DB. Experimental results show that T2VQA outperforms existing T2V generation metrics and state-of-the-art VQA models, validating its effectiveness in measuring the perceptual quality of text-generated videos. Fig. 1 shows the overview of the proposed T2VQA-DB and T2VQA.

We summarize our contributions as follows:

- (1) We establish the T2V dataset with *the largest scale to date*, named T2VQA-DB, which includes 10,000 text-generated video sequences and their corresponding MOSs gained from 27 subjects.
- (2) We propose a novel transformer-based model for better evaluating the quality of text-generated videos, named T2VQA. The model dissolves the problem from *text-video alignment* and *video fidelity* perspectives, and then it leverages the ability of an LLM to give a subjective-aligned prediction of the video quality.
- (3) The proposed T2VQA outperforms existing T2V generation metrics and SOTA VQA models on T2VQA-DB and other T2V datasets, indicating the effectiveness of T2VQA. Qualitative experiments show that T2VQA can benefit in measuring the performance of T2V generation algorithms, giving it practical application prospects.

2 Related Works

2.1 Text-to-video Dataset

To the best of our knowledge, only a few T2V datasets have been proposed. They mainly have the following two issues: (1) Insufficient scale: Chivileva *et al.* [5] propose a dataset with 1,005 videos generated by 5 T2V models following [27]. EvalCrafter [30] builds ECTV using 700 prompts and 8 T2V models, resulting in 5,600 videos in total. Similarly, FETV [31] is composed of 2,476 videos generated by 619 prompts, and 4 T2V models. VBench [17] has a larger scale with in total of ~1.7k prompts and 4 T2V models. TVGE [57] uses the prompts from ECTV and uses 5 T2V model for video generation. Such scales are not sufficient for the training of deep learning-based models, and cannot represent current T2V algorithms. (2) Limited human annotation: ITU-standard [18] requires at least 15 human annotators for subjective study. The dataset proposed by Chivileva *et al.* is the only one that meets the standard. ECTV and FETV only have 3 users for annotation. VBench [17] does not specify the number of human annotators.

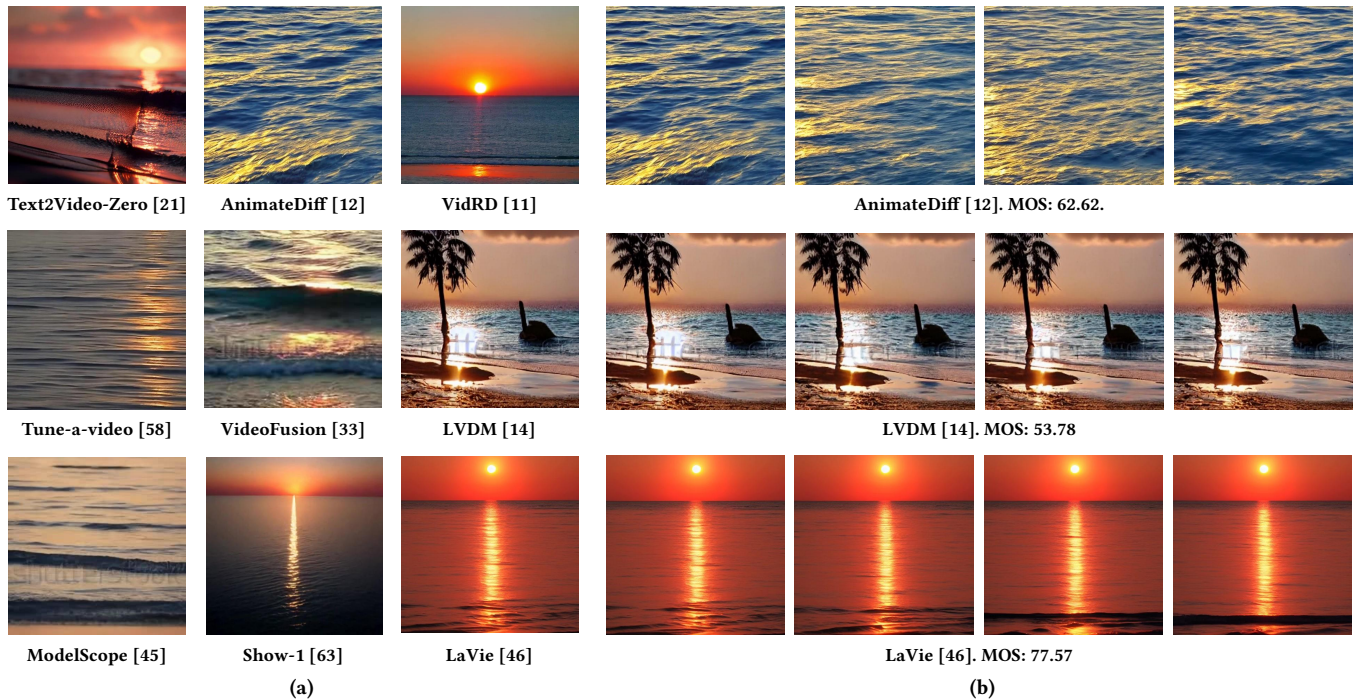


Figure 2: Video examples generated by prompt: *Sunset on the sea*. (a) Overview of video frames generated by 9 models. (b) Videos generated by AnimatedDiff [12], LVDM [14], LaVie [46], and their MOSs.

2.2 Metrics for T2V Generation

IS [38] and FVD [44] are the two most commonly used metrics for evaluating the quality of generated videos. IS uses the Inception feature to present both image/video quality. FVD measures the distance between the generated video and the natural video. However, both metrics are criticized for poor correlation with human visual perception. CLIPSim [48] measures the text-video alignment by using CLIP [34]. After measuring the similarity between the text and each video frame, it averages them to get the final score. As a result, it only evaluates videos from the image level, losing the information from the temporal domain. As well as it doesn't consider the video quality.

Though many works targeting the evaluation of text-generated images have been proposed [22, 25–27, 59–61, 65], there are only a few metrics tailored for text-generated video evaluation have been proposed. Among them, ViCLIP [47] is a CLIP-based metric for measuring text-video alignment. Chivileva *et al.* [5] proposes an ensemble video quality metric that integrates text similarity and naturalness. EvalCrafter [30] and VBench [17] build benchmarks to evaluate text-generated video from 18 and 16 objective metrics respectively. FETV [31] and T2V-Score [57] both propose separate metrics for text-video alignment and video quality, without an overall perceptual score for text-generated videos. There is still lack of a simple and effective metric to evaluate the quality of text-generated videos.

Besides the aforementioned metrics, VQA models can also be used for the evaluation of text-generated videos. BVQA [24] transfers knowledge from Image Quality Assessment (IQA) databases

and then trains on the target VQA database, *e.g.*, KoNVid-1k[16]. SimpleVQA [41] extracts spatial and motion features to regress to the final score. FAST-VQA [49] proposes “fragments” as a novel sampling strategy and the fragment attention network to accommodate fragments as inputs. [8, 9, 23, 66, 69, 72, 73] are works for the evaluation of enhanced videos and digital humans. DOVER [52] proposes to view the quality assessment problem from technical and aesthetic perspectives, while BVQI [50] and MaxVQA [51] integrate text prompts (*e.g.*, good, bad) into VQA. With the development of Multi-modal Large Language Models (MLLMs), researchers have started to leverage MLLMs to solve VQA problems, as they have been trained on massive data. Q-Bench [53] proves that MLLMs can address preliminary low-level visual tasks. Q-Align [55] proposes to train a MLLM using text-defined levels (*e.g.*, fine, poor) and achieves SOTA results on IQA and VQA tasks. Though many VQA models have been proposed, they are originally designed for natural videos and do not consider text-video alignment.

2.3 Text-to-video Generation

T2V generation refers to a form of conditional video generation, where text descriptions are used as conditioning inputs to generate high-fidelity videos. A common practice is to extend pre-trained Text-to-Image (T2I) models with temporal modules. CogVideo [15] is based on CogView2 [7] and proposes a multi-frame-rate hierarchical training strategy to better align text-video clips. Make-a-video [39] adds effective spatial-temporal modules on a diffusion-based T2I model (*i.e.*, DALLE-2 [35]). VideoFusion [33] also leverages the DALLE-2 and presents a decomposed diffusion process.

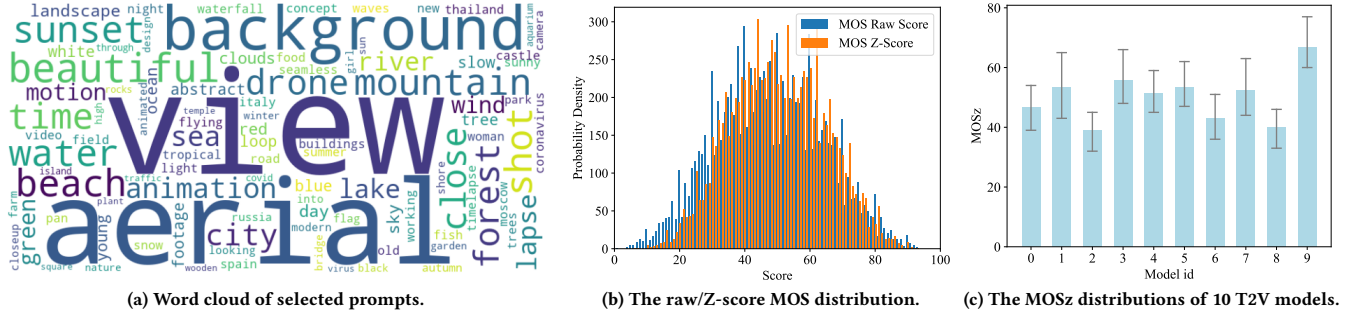


Figure 3: (a) The word cloud of the prompts used in T2VQA-DB. (b) The distribution of the raw/Z-score MOS. (c) Z-score MOS distributions of 10 T2V models. The model IDs represent sequentially Text2Video-Zero, AnimateDiff, Tune-a-video⁽¹⁾, VidRD, ModelScope, VideoFusion, LVDM, Show-1, Tune-a-video⁽²⁾, LaVie.

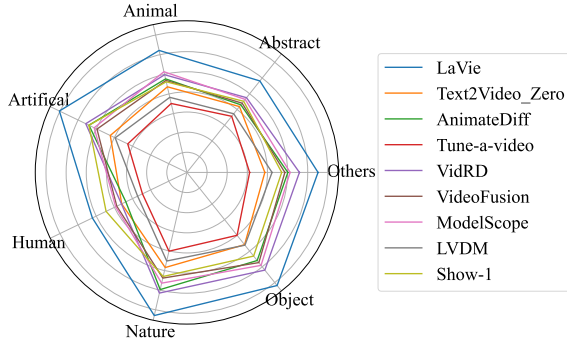


Figure 4: Comparison of models performance on different prompt types.

LVDM [14], Text2Video-Zero [21], Tune-A-Video [58], Animate-Diff [12], Video LDM [2], MagicVideo[71], ModelScope [45], and VidRD [11] are models that inherit the success of Stable Diffusion (SD) [37] for video generation. Show-1 [63] integrates both pixel-based and latent-based text-to-Video Diffusion Models (VDMs). LaVie [46] extends the original transformer block in SD to a spatio-temporal transformer. Recently, OpenAI releases Sora [3], a T2V model that is capable of generating 60s high-fidelity videos, considered as a game changer in T2V generation.

3 Subjective-Aligned Text-to-Video Dataset

Tab. 1 shows most existing T2V datasets have relatively small number of videos, which is not sufficient to represent the diverse performance of T2V generation models. Consequently, we propose a Text-to-Video Quality Assessment DataBase, named T2VQA-DB, including 10,000 videos generated by 9 different T2V models. 1,000 prompts are used and 27 subjects are invited to obtain the MOS of each video. In this section, we will describe the establishment of T2VQA-DB and the subjective experiment.

3.1 Data Preparation

3.1.1 Prompt Selection. To guarantee the diversity of the dataset, the prompts used for T2V generation should cover as many aspects as possible. Following [60], we use the same graph-based algorithm from [40] for prompt selection. We first randomly sample 1 million

prompts from WebVid-10M [1], which contains 10 million video-text pairs scraped from the stock footage sites. Each prompt is encoded to a vector representation by Sentence-BERT [36]. The graph-based algorithm integrates them into k groups according to cosine distance. k is a hyper-parameter and we set $k = 100$, resulting roughly 10,000 prompts in each group. Finally, we randomly sample 10 prompts in each group, forming the 1,000 prompts in T2VQA-DB. Fig. 3a shows the word cloud of the collected prompts.

3.1.2 Video Generation. We use in total 9 different models for video generation, including Text2Video-Zero [21], AnimateDiff [12], Tune-a-video [58], VidRD [11], VideoFusion [33], ModelScope [45], LVDM [14], Show-1 [63], and LaVie [46]. For Tune-a-video, we utilize two different pre-trained weights, resulting in a total of 10 models for generation. Compared to other T2V datasets, we utilize current advanced T2V generation models as much as possible, making T2VQA-DB more representative. Since the default resolution, video length, and frame rate are different in each model, we unify the video format as 512×512 , 16 frames, and 4fps.

3.2 Subjective Study

To obtain the MOS of each video, we invite 27 subjects to score the perceptual quality of each video. The subjects are asked to score mainly from two aspects, in terms of text-video alignment and video fidelity. Text-video alignment refers to how the generated video content matches the text description. Video fidelity refers to degrees of distortion, saturation, motion consistency, and content rationality. The subjects use a slider ranging from 0 to 100 to give the final score of each video. After having the raw MOS of each subject, we conduct normalization to avoid inter-subject scoring differences as Z-score MOS (MOSz). That is:

$$MOSz_i = \frac{1}{N} \sum_{j=1}^N Res\left(\frac{r_{ij} - \mu_j}{\sigma_j}\right). \quad (1)$$

i and j refer to the index of videos and subjects. r is the raw score, and $\mu_j = \frac{1}{M} \sum_{i=1}^M r_{ij}$, $\sigma_j = \sqrt{\frac{1}{M-1} \sum_{i=1}^M (r_{ij} - \mu_j)^2}$. M is the number of videos scored by each subject. N is the number of scores on one video. $Res(\cdot)$ is the rescaling function, converting the distribution of Z-scores into a mean of 50 and a standard deviation of 16.6. Fig. 3b shows the distributions of raw MOS and Z-score MOS.

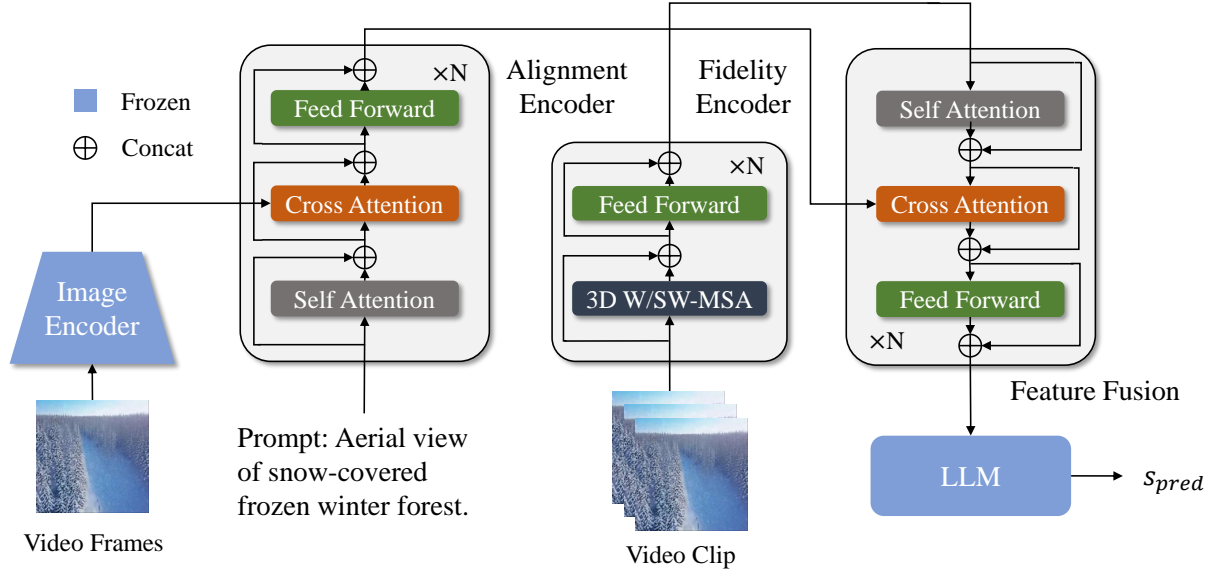


Figure 5: Overview framework of T2VQA. Features from text-video alignment and video fidelity perspectives are extracted. After a cross-attention based fusion, an LLM is utilized for regression.

3.3 Dataset Analysis

We conduct comprehensive experiments and analysis on T2VQA-DB. We first investigate each model’s performance in T2VQA-DB. The visualization of the Z-score MOS distributions of 10 models is shown in Fig. 3c. LaVie has the highest average MOS of 66.9, while two Tune-a-Video models have the lowest of 39.1 and 39.9. The reason for the poor performance of Tune-a-Video is mainly the low inter-frame consistency, as shown in Fig. 7c.

Based on the prompt contents, we classify the collected prompts into 6 categories, including nature, human, artificial, animal, object, and abstract. The ones that cannot be categorized into the 6 classes are labeled as “others”. After classification, we have 327 prompts for nature, 119 for human, 196 for artificial, 121 for animal, 66 for object, 135 for abstract, and 36 for others. Subsequently, we compare the models’ performance over different prompt types. As shown in Fig. 4, LaVie outperforms the other models on all types of prompts. Tune-a-video has the worst performance, which is consistent with the analysis in Fig. 3c. LVDM has the second-worst performance except in the “others” type. The performance of the rest models has negligible differences. Fig. 4 also shows prompts labeled as human have the worst performance in all models. The reason could be that human faces and actions require more sophisticated modeling compared with other categories. Some models also show a preference for object and nature types of prompts to a small extent.

4 Subjective-Aligned T2V Metric

Based on T2VQA-DB, we propose a novel model that leverages transformer-based architecture for Text-to-video Quality Assessment (T2VQA). The model dissolves the task into two perspectives, in terms of text-video alignment and video fidelity. After feature extraction and feature fusion, an LLM is used for quality regression. Fig. 5 shows the overview of the architecture of T2VQA. We will introduce the detailed design of T2VQA below.

4.1 Text-video Alignment Encoder

Text-video alignment refers to the conformity between the video content and the text description. CLIP [34] and BLIP [29] have strong abilities for zero-shot text-image matching. [48] proposes CLIPSim, which uses CLIP to calculate the similarities between text and each frame of the video and then take the average value. However, former works [5, 17, 30, 31] and results in Tab. 2 show that simply using CLIP or BLIP has a low correlation with the authentic subjective scores. Following works of MLLMs [6, 28, 62], we use the pre-trained BLIP image encoder as the video frame encoder. We freeze the weights of the image encoder that encodes each frame separately. We use the BLIP text encoder as the alignment encoder. The alignment encoder takes the encoded text and video frame as inputs. They interact through the cross-attention module and the encoder eventually outputs a feature representing the text-frame similarity. We concatenate features from each text-frame pair. Given a set of N video frames $\{v_i \in \mathbb{R}^{3 \times H \times W}\}_{i=1}^N$ and the text prompt t , we have:

$$f_b = \text{cat}(\{\text{BLIP}(t, v_i)\}_{i=1}^N), \quad (2)$$

where *cat* is short for concatenation.

4.2 Video Fidelity Encoder

Video fidelity refers to the perception of distortion from spatial and temporal domains. In the spatial domain, common distortion types include blurriness, noises, low/high contrast, *etc.* Temporal distortions include jitter, stall, motion blur, *etc.* In this perspective, the task can be seen as a common VQA task. Swin-T [32] has been proven for its excellent ability in various VQA tasks [49, 52]. By using 3D-shifted window-based multi-head self-attention (SW-MSA), Swin-T has a strong ability to analyze videos from spatial and temporal domains. Therefore, we utilize Swin-T as the backbone of

Table 2: Performance of the SOTA models and T2VQA. The best model is highlighted in each column. [Bold: the best].

Type	Models	T2VQA-DB Validation				ECTV Testing		TVGE Testing		Sora Testing	
		SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑
zero-shot	CLIPSim [34]	0.1047	0.1277	0.0702	21.683	0.3111	0.3175	0.2765	0.3013	0.2116	0.1538
	BLIP [29]	0.1659	0.1860	0.1112	18.373	0.3841	0.3877	0.3437	0.3696	0.2126	0.1038
	ImageReward [60]	0.1875	0.2121	0.1266	18.243	0.5192	0.5107	0.4740	0.4879	0.0992	0.0415
	ViCLIP [47]	0.1162	0.1449	0.0781	21.655	0.4130	0.4105	0.3429	0.3691	0.2567	0.1844
	UMTScore [31]	0.0676	0.0721	0.0453	22.559	0.2406	0.2214	0.2155	0.2199	0.2594	0.0840
finetuned	SimpleVQA [41]	0.6275	0.6338	0.4466	11.163	0.1959	0.1391	0.3297	0.2219	0.0340	0.2344
	BVQA [24]	0.7390	0.7486	0.5487	15.645	0.2929	0.2829	0.4283	0.4229	0.4235	0.2489
	FAST-VQA [49]	0.7173	0.7295	0.5303	10.595	0.3101	0.3246	0.3335	0.3466	0.4301	0.2369
	DOVER [52]	0.7609	0.7693	0.5704	9.8072	0.3118	0.3289	0.3943	0.4058	0.4421	0.2689
	Q-Align [55]	0.7601	0.7768	0.5860	10.9911	0.4356	0.4401	0.5317	0.5209	0.3109	0.1997
Ours	T2VQA	0.7965	0.8066	0.6058	9.0221	0.5694	0.5823	0.6898	0.6912	0.6485	0.3124

the fidelity encoder to extract features that represent video fidelity. Given a video clip $v \in \mathbb{R}^{3 \times N \times H \times W}$, we have:

$$f_s = \text{SWIN}(v). \quad (3)$$

4.3 Feature Fusion

Inspired by BLIP-2 [28] and InstructBLIP [6], after having the features from perspectives of text-video alignment and video fidelity, we design a transformer-based fusion module to fuse those two features. The module includes N blocks with self-attention, cross-attention, and feed-forward layers in each block. The fidelity feature f_s first goes through self-attention layers, and then it interacts with the alignment feature f_b in cross-attention layers (and every other transformer block). The fusion module helps the model to unify the features from two perspectives to have a more comprehensive understanding of the video characteristics. We initialize the fusion module using BERT_{base} [20].

4.4 Quality Regression

LLMs have been proven to be competitive on quality assessment tasks [54–56, 67, 68]. Inspired by them, we also utilize an LLM as the quality regression module in T2VQA. We first design a text instruction prompt, i.e., “Please rate the quality of this video.”, to guide the LLM. The encoded instruction and the fused feature are concatenated as the input of the LLM. Following [55, 64], we supervise the LLM to output one among five ITU-standard [18] levels (*bad*, *poor*, *fair*, *good*, and *excellent*) to represent the quality of the videos, denoted as <level>. We assign them weights of 1–5 in order. Since the logit at <level> in LLM is the probability distribution of all tokens, it can be used to represent how the LLM predicts the quality of the video. Therefore, a softmax for each token is calculated and multiplied by its weight. We have the final predicted score as:

$$s_{pred} = \sum_{i=1}^5 i \times \text{softmax}(\lambda_i) = \sum_{i=1}^5 i \times \frac{e^{\lambda_i}}{\sum_{j=1}^5 e^{\lambda_j}}, \quad (4)$$

where λ_i is the probability distribution of the i -th <level> token.

5 Experiments and Results

5.1 Implement Details

5.1.1 Train-test splitting. When training and testing on T2VQA-DB, we follow the common practice of dataset splitting by leaving out 80% for training, and 20% for testing. To eliminate the bias in one single split, we randomly split the dataset 10 times, and use the average results for performance comparison.

5.1.2 Training Settings. We utilize the large model of BLIP for the extraction of text-video alignment feature. We initialize the fidelity encoder using Swin-T pre-trained on Kinetics-400 [19] dataset. For the LLM, we use the 7B model of Vicuna v1.5 [70], which is fine-tuned from Llama 2 [43]. It is worth noting that both BLIP image encoder and LLM are frozen during training.

During training and testing, we first uniformly sample 8 frames out of an input video, and then we resize them to 224×224 . We use Adam optimizer initialized by a learning rate of $1e-5$. The learning rate decays under a cosine scheduler from 1 to 0. We train T2VQA for 30 epochs under a batch size of 4 on a server with one NVIDIA GeForce RTX 4090.

5.1.3 Loss Function. Following [49], we use differentiable Pearson Linear Correlation Coefficient (PLCC) and rank loss as loss functions. PLCC is a common criterion used for evaluating the correlation between sequences, while rank loss is introduced to help the model distinguish the relative quality of videos better. The final loss function is defined as:

$$L = L_{plcc} + \lambda \cdot L_{rank}. \quad (5)$$

λ is a hyper-parameter for balancing, and is set to 0.3 in training.

5.1.4 Evaluation Metrics. Along with PLCC, we include Spearman’s Rank-order Correlation Coefficient (SRCC), Kendall’s Rank-order Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE) as performance criteria. Better models should have larger SRCC, KRCC, and PLCC scores, but conversely for RMSE. Before calculating PLCC, we follow [10] to map the objective score to the subject score using a four-parameter logistic function.

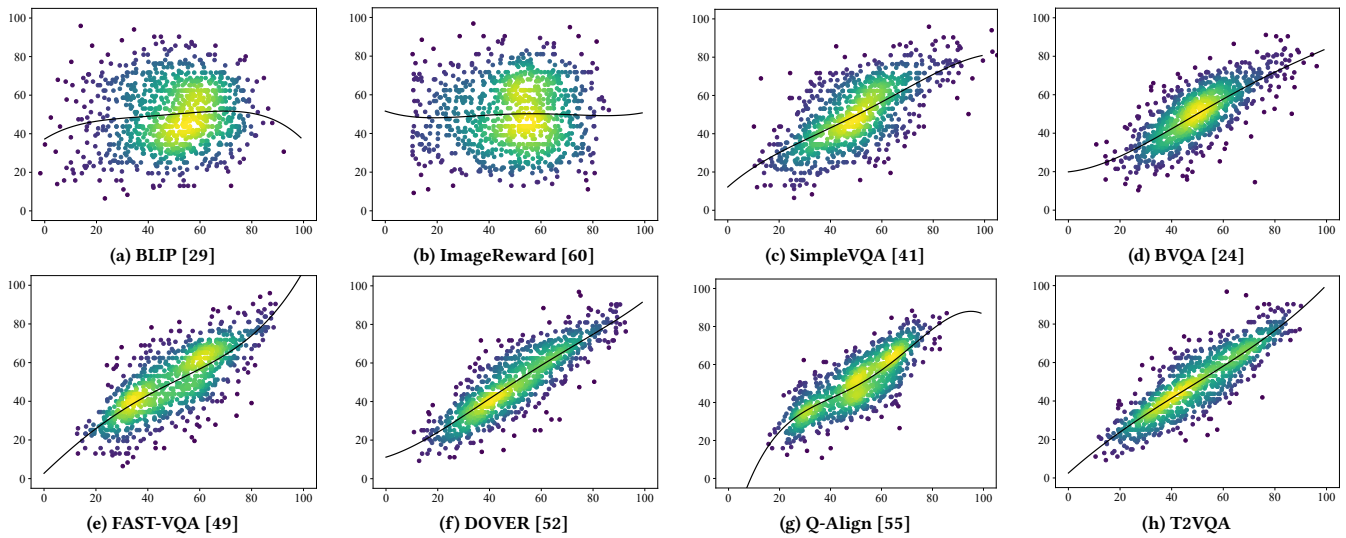


Figure 6: Scatter plots of the predicted scores vs. MOSs. The curves are obtained by a four-order polynomial nonlinear fitting. The brightness of scatter points from dark to bright means density from low to high.

5.2 Performance Comparison

5.2.1 Reference Algorithms. We use CLIPSim [48], BLIP [29], ImageReward [60], ViCLIP [47], UMTScore [31], SimpleVQA [41], BVQA [24], FAST-VQA [49], DOVER [52], and Q-Align [55] as the reference algorithms. Q-Align is considered the SOTA VQA method to date. CLIPSim averages the similarity values between the text and each video frame. We adopt the same operation on BLIP and ImageReward to tune IQA metrics into VQA metrics. ViCLIP and UMTScore are metrics designed for measuring text-video alignment. SimpleVQA, BVQA, FAST-VQA, DOVER, and Q-Align are models designed for general VQA tasks. We use the pre-trained weights of CLIPSim, BLIP, and ImageReward for zero-shot testing, since they are originally designed for generated content. SimpleVQA, BVQA, FAST-VQA, DOVER, and Q-Align are originally designed for UGC videos, so we finetune them on T2VQA-DB. All results are averaged after ten-fold splitting.

5.2.2 Results Analysis. The first four columns of number in Tab. 2 show the performance comparison between T2VQA and other SOTA models on T2VQA-DB. Results show that T2VQA performs best in SRCC, surpassing Q-Align by 4.79% in SRCC and 3.84% in PLCC. The zero-shot models all have relatively low scores. They either only consider the text-video alignment or don’t analyze the temporal domain information within video frames. The VQA models have higher scores, indicating that video fidelity heavily affects the assessment of text-generated video quality. However, a single perspective from video fidelity cannot address the problem properly, as there are circumstances where a high-fidelity video is generated but does not match the prompt.

Fig. 6 shows scatter plots between the predicted scores and the Z-Score MOSs of BLIP [29], ImageReward [60], ViCLIP [47], SimpleVQA [41], BVQA [24], FAST-VQA [49], DOVER [52], and T2VQA. The figure uses 1,000 videos, which are randomly sampled from the testing set in one split of T2VQA-DB. A better model should have

a fitted curve close to the diagonal and have less dispersed scatter points. As shown in Fig. 6, T2VQA also outperforms the others.

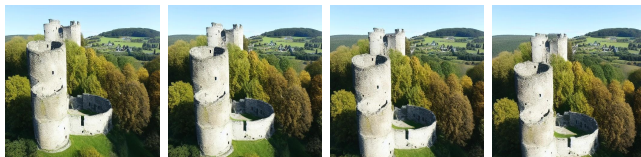
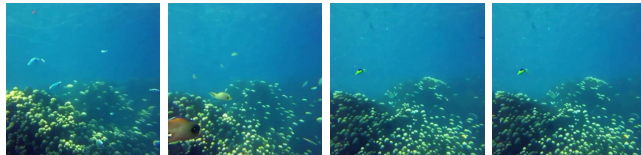
5.2.3 Cross-dataset Validation. ECTV [30] and TVGE [57] are the only two T2V datasets that have released their subjective scores. We randomly sample 800 videos from each dataset for cross-dataset validation. Besides, Sora [3] has been considered the SOTA T2V generation model. We collect the videos generated by Sora from its official website to validate the generalization of T2VQA and other models on high-fidelity videos. We invite 20 annotators to score the quality of each Sora video. T2VQA and other reference VQA models are trained on T2VQA-DB and tested on the sampled ECTV, TVGE, and Sora videos. We report the SRCC and PLCC between the models predictions and the ground truth MOSs. The results are listed in the six columns on the right side of Tab. 2.

Experimental results show that T2VQA has the best generalization ability among all models. Noticed that there is a performance drop between the validation on T2VQA-DB and the testing on other datasets. That’s because videos from the other datasets have the attributes of high resolution, high frame rate, and long length, which current open-sourced T2V models are not able to generate. We will include videos with high fidelity in T2VQA-DB in future work.

5.2.4 Qualitative Analysis. We also conduct a qualitative analysis on three examples with good, fair, and poor quality. We use SimpleVQA [41], BVQA [24], FAST-VQA [49], DOVER [52], and T2VQA to predict their quality. Fig. 7 presents the prompts, video frames, and model predictions on the three examples, and Tab. 3 lists the models’ predictions and MOSs. Results show that T2VQA has more subjective-aligned predictions. Fig. 7a shows an example with a relatively high score. The scene in the video matches the description in the prompt well, and the video frames are clear and consistent. Fig. 7b is a medium-level example. Though the video matches the prompt basically, the video frames suffer from blurriness, which is reflected in the MOS and T2VQA’s prediction. Fig. 7c shows the

Table 3: T2VQA and other models predictions on 3 example videos. Colored numbers represent the distance between predictions and the ground truths. Red: the best.

Model	Prompt 1	Prompt 2	Prompt 3
SimpleVQA [41]	72.89 _{-5.33}	61.59 _{+3.88}	49.95 _{+17.39}
BVQA [24]	73.12 _{-5.1}	60.37 _{+2.66}	45.68 _{+13.12}
FAST-VQA [49]	85.28 _{+7.06}	44.19 _{-13.52}	38.91 _{+6.35}
DOVER [52]	89.11 _{+10.89}	51.93 _{-5.78}	36.66 _{+4.1}
T2VQA(Ours)	81.61 _{+3.39}	57.11 _{-0.6}	29.07 _{-3.49}
MOS (gt)	78.22	57.71	32.56

**(a) Prompt: Castle ruins on the hill in the middle of a beautiful landscape.****(b) Prompt: Underwater world with different fishes, corals, and stones.****(c) Prompt: In a hot cast-iron cauldron, the cook pours oil to fry the meat (liver).****Figure 7: Three videos with good, fair, and poor quality.**

worst case. The video fails to accurately present the description in the prompt. Besides, it loses consistency between video frames. Although it has a high definition in each frame separately, it still has low scores in both MOS and T2VQA’s prediction.

5.3 Ablation Studies

To validate the effectiveness of each module in T2VQA, we conduct thorough ablation studies, including the alignment and fidelity encoder, the fusion module, and the regression module. Results are listed in Tab. 4. All results are averaged after 10-fold splitting.

5.3.1 Alignment Encoder. T2VQA utilizes the BLIP image encoder and text encoder as the alignment encoder. CLIP is another widely used text-image encoder. It has a similar structure and ability to measure text-video alignment as BLIP. We replace the BLIP image and text encoder in T2VQA with CLIP to determine which one has the better performance.

Experimental results show that using CLIP as the alignment encoder suffers from severe performance degradation. The reason could be that the image and the text are encoded separately in CLIP. While in BLIP, the encoded image and the text features interact in the cross-attention module.

Table 4: Results of ablation studies. [Keys: Bold: the best].

Models	Validation			
	SRCC \uparrow	PLCC \uparrow	KRCC \uparrow	RMSE \downarrow
T2VQA-CLIP	0.7296	0.7347	0.5385	10.5141
T2VQA-resnet	0.7610	0.7730	0.5715	9.8152
T2VQA-cat	0.7734	0.7854	0.5839	9.4034
T2VQA-linear	0.7808	0.7919	0.5891	9.3011
T2VQA-nonlinear	0.7850	0.7983	0.5954	9.1755
T2VQA(Ours)	0.7965	0.8066	0.6058	9.0221

5.3.2 Fidelity Encoder. In T2VQA, we use the Swin-T as the backbone of the fidelity encoder. To investigate the effectiveness of the transformer-based architecture, we conduct the control experiment by using the convolution-based 3D ResNet [13] as the fidelity encoder. The results show that Swin-T has a superior performance to ResNet, validating its effectiveness.

5.3.3 Fusion and Regression Modules. Besides, we compare our cross-attention fusion strategy with the simple concatenation fusion. The latter is the simplest yet most commonly used fusion strategy. In T2VQA, we take advantage of the strong ability of an LLM for quality regression. We test the commonly used linear regression and non-linear regression to compare with the LLM regression in T2VQA. For linear regression, we use two full-connected layers with 128 neurons in the first layer and 1 neuron in the second. For non-linear regression, we use two 1D convolution blocks with kernel size set to 1. We also set the channel number to 128 in the first block and 1 in the second.

Results in Tab. 4 show that T2VQA achieves the best performance in all evaluating metrics, indicating its effectiveness. The models using concatenation, linear regression, and non-linear regression have similar performance, yet they are all inferior to T2VQA, indicating that the cross-attention fusion and LLM have achieved non-negligible improvement to the model.

6 Conclusion

In conclusion, in this paper, we are dedicated to giving a subjective-aligned prediction of the quality of a text-generated video. For that purpose, we establish a T2V dataset with the largest scale, named T2VQA-DB. The dataset includes 10,000 videos generated by 9 advanced T2V models. We also conduct a subjective study to obtain the MOSs on the overall video quality. Based on T2VQA-DB, we propose a novel transformer-based model for text-to-video quality assessment, named T2VQA. The model extracts features of a video from text-video alignment and video fidelity perspectives respectively. After fusing the features, an LLM is utilized to regress the final prediction. The experimental results indicate that T2VQA is effective in evaluating the quality of text-generated videos.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62301310, 62225112, and U1908210, and in part by the Shanghai Pujiang Program under Grant 22PJ1406800, and in part by Sichuan Science and Technology Program under Grant 2024NSFSC1426.

References

- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision*.
- [2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.
- [3] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). <https://openai.com/research/video-generation-models-as-world-simulators>
- [4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [5] Iya Chivileva, Philip Lynch, Tomas E Ward, and Alan F Smeaton. 2023. Measuring the Quality of Text-to-Video Model Outputs: Metrics and Dataset. *arXiv preprint arXiv:2309.08009* (2023).
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv:2305.06500* [cs.CV]
- [7] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems* 35 (2022), 16890–16902.
- [8] Yunlong Dong, Xiaohong Liu, Yixuan Gao, Xunchu Zhou, Tao Tan, and Guangtao Zhai. 2023. Light-vqa: A multi-dimensional quality assessment model for low-light video enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1088–1097.
- [9] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2023. Vdpve: Vqa dataset for perceptual video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1474–1483.
- [10] Video Quality Experts Group et al. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment. In *VQEG meeting, Ottawa, Canada, March, 2000*.
- [11] Jiayi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. 2023. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549* (2023).
- [12] 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* (2023).
- [13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6546–6555.
- [14] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221* (2022).
- [15] 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* (2022).
- [16] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.
- [17] 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. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [18] B. T. ITU-R. 2002. Methodology for the subjective assessment of the quality of television pictures. <https://www.itu.int/rec/R-REC-BT.500>.
- [19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).
- [20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-HLT*, Vol. 1. 2.
- [21] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15954–15964.
- [22] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).
- [23] Tengchuan Kou, Xiaohong Liu, Wei Sun, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Ning Liu. 2023. Stablvqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1066–1076.
- [24] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. 2022. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 9 (2022), 5944–5958.
- [25] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. 2024. Aigiq-20k: A large database for ai-generated image quality assessment. *arXiv preprint arXiv:2404.03407* 2, 3 (2024), 5.
- [26] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. 2024. Q-Refine: A Perceptual Quality Refiner for AI-Generated Image. *arXiv preprint arXiv:2401.01117* (2024).
- [27] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology* (2023). <https://doi.org/10.1109/TCSVT.2023.3319020>
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- [29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.
- [30] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejiong Zeng, Raymond Chan, and Ying Shan. 2023. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440* (2023).
- [31] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems* 36 (2024).
- [32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3202–3211.
- [33] Zhengxiang Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10209–10218.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- [36] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
- [37] 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*. 10684–10695.
- [38] 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 (2016).
- [39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiuyan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
- [40] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975* (2022).
- [41] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. 2022. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*. 856–865.
- [42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
- [43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutit Bhoale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

- [44] 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* (2018).
- [45] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscape text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).
- [46] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yanan He, Jiashuo Yu, Peiqing Yang, et al. 2023. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103* (2023).
- [47] Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942* (2023).
- [48] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806* (2021).
- [49] 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*. Springer, 538–554.
- [50] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Exploring Opinion-unaware Video Quality Assessment with Semantic Affinity Criterion. In *International Conference on Multimedia and Expo (ICME)*.
- [51] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Towards Explainable Video Quality Assessment: A Database and a Language-Prompted Approach. In *ACM MM*.
- [52] 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 *International Conference on Computer Vision (ICCV)*.
- [53] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181* (2023).
- [54] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024. Q-Instruct: Improving Low-level Visual Abilities for Multi-modality Foundation Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [55] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2023. Q-align: Teaching llms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* (2023).
- [56] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. 2024. Towards Open-ended Visual Quality Comparison. *arXiv preprint arXiv:2402.16641* (2024).
- [57] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. 2024. Towards A Better Metric for Text-to-Video Generation. *arXiv:2401.07781* [cs.CV]
- [58] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.
- [59] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2096–2105.
- [60] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).
- [61] Liu Yang, Huiyu Duan, Long Teng, Yucheng Zhu, Xiaohong Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. 2024. Aigcoiq2024: Perceptual quality assessment of ai generated omnidirectional images. *arXiv preprint arXiv:2404.01024* (2024).
- [62] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257* (2023).
- [63] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. 2023. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818* (2023).
- [64] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind Image Quality Assessment via Vision-Language Correspondence: A Multi-task Learning Perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [65] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2023. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 440–445.
- [66] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. Advancing Zero-Shot Digital Human Quality Assessment through Text-Prompted Evaluation. *arXiv preprint arXiv:2307.02808* (2023).
- [67] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al. 2023. Q-Boost: On Visual Quality Assessment Ability of Low-level Multi-Modality Foundation Models. *arXiv preprint arXiv:2312.15300* (2023).
- [68] Zicheng Zhang, Haoning Wu, Yingjie Zhou, Chunyi Li, Wei Sun, Chaofeng Chen, Xiongkuo Min, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. 2024. LMM-PCQA: Assisting Point Cloud Quality Assessment with LMM. *arXiv preprint arXiv:2404.18203* (2024).
- [69] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Kang Fu, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2024. A Reduced-Reference Quality Assessment Metric for Textured Mesh Digital Humans. In *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2965–2969.
- [70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).
- [71] 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* (2022).
- [72] Xunchu Zhou, Xiaohong Liu, Yunlong Dong, Tengchuan Kou, Yixuan Gao, Zicheng Zhang, Chunyi Li, Haoning Wu, and Guangtao Zhai. 2024. Light-VQA+: A Video Quality Assessment Model for Exposure Correction with Vision-Language Guidance. *arXiv preprint arXiv:2405.03333* (2024).
- [73] Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Zhihua Wang, Xiao-Ping Zhang, and Guangtao Zhai. 2024. THQA: A Perceptual Quality Assessment Database for Talking Heads. *arXiv preprint arXiv:2404.09003* (2024).