CleveReward: Contrastive Learning-Engined Video Reward Training on Different Benchmark Datasets

Yuanming Yang[1] **Xiaoqian Liu**[2] **Jian Chen**[2]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Department of Automation, Tsinghua University
{yym22,lxq21,chenjian20}@mails.tsinghua.edu.cn

Abstract

With the rapid development of text-to-video (T2V) generation models, the ability to synthesize high-quality videos from textual descriptions has significantly improved. However, despite the continuous enhancement of video quality, the consistency between the generated video and the text remains a major challenge. In particular, inconsistencies or hallucinations between the generated video and the text can negatively impact the output quality. Therefore, accurately assessing the instruction-following capability of video generation models is crucial. Given the high cost and limited scalability of manual evaluation, this paper proposes an end-to-end reward model to automate the evaluation of instruction-following performance in video generation models. A key issue in training video understanding models and reward models is data insufficiency. The inconsistency in labeling standards and dimensions across video scoring datasets complicates dataset integration. To address this problem, we introduce a contrastive learning-based video reward model, **CleveReward**, which converts video scoring datasets such as T2VQA and videofeedback into pairwise formats and trains using contrastive learning. Experimental results demonstrate that CleveReward effectively trains across datasets and holds the potential to surpass current state-of-the-art video reward models.Furthermore, we introduce **VideoCross**, an open-source dataset designed to support contrastive learning. VideoCross integrates data from various standards, reduces redundancy, and enhances consistency, providing high-quality data support for model training. By constructing positive and negative sample pairs, VideoCross helps improve the model's understanding of the alignment between text and video, thereby enhancing the instruction-following performance of video generation models. During training, we employed two advanced video understanding models, Qwen2-vl-7b-chat and cogvlm2-video, and utilized 16 A800 GPUs for hardware acceleration.This research offers new insights into the evaluation and optimization of text-to-video generation models and advances the development of reward models and datasets to better support the application and advancement of video generation technology.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

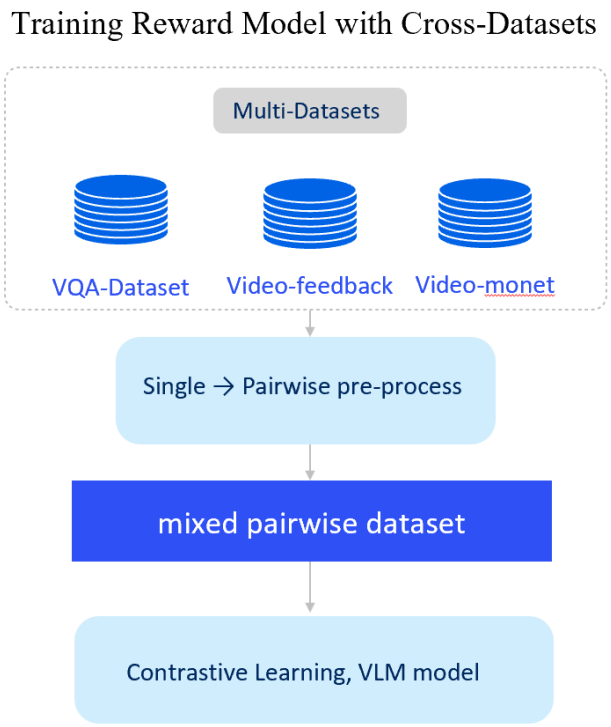

Figure 1: Training Scheme of CleveReward

1 Introduction

Text-to-video (T2V) generation models have made significant strides in recent years, enabling the synthesis of high-quality videos from textual descriptions.

These applications and products show great promise in producing high-quality, longer-duration videos that adhere to physical laws.

In T2V research, automating the evaluation of AI-generated videos is a crucial area of study. Not only can it save on the costs and time associated with manual evaluation, but it also allows for the optimization of video generation models using reinforcement learning algorithms, such as DPO and PPO, based on reward models..

Therefore, accurately assessing the instruction-following performance of video generation models is crucial. Given the high cost and limited scalability of manual evaluations, we propose an end-to-end reward model to automate the evaluation of this capability in text-generated videos.

As video understanding models continue to improve in their general capabilities, fine-tuning video understanding models to develop reward models has become more prevalent. While both training video understanding models and reward models involve similar architectures, there is a significant disparity in the amount of data available. The pre-training phase for video understanding models typically involves around 1.5 billion images or video-text pairs, whereas training reward models usually only has about 30,000 labeled pairs. Although supervised fine-tuning does not require datasets as extensive as those used in pre-training, the limited amount of data undeniably reduces the accuracy of the reward models.

One major reason for data insufficiency is that video-text annotated datasets are readily shareable. Researchers often utilize and seamlessly merge previously labeled datasets with their new annotations. However, video-score pairs are marked by disparate dimensions of interest, varying score ranges, and evaluation standards used by different researchers, preventing the integration of these datasets. Consequently, research on video reward models frequently discards prior annotated datasets and

instead starts from scratch according to their own annotation rules. This significantly decreases the efficiency of dataset usage.

To address this issue, we propose the first contrastive learning-based video reward model: **CleveReward** (Contrastive Learning Video Reward). Initially, we amalgamated video scoring datasets such as T2VQA and videofeedback, converting them into a pairwise format. We then employed contrastive learning for training. Our findings indicate that utilizing contrastive learning allows effective training across datasets with different scoring systems. Although our model is still under training, it holds promise to surpass the current state-of-the-art video reward models.

2 Related Work

2.1 Text-to-Video Generative Models

The development of Text-to-Video (T2V) generative models that incorporate physical laws has been advancing rapidly. Based on Text-to-Video (T2V) generation models, many applications and products have also emerged, such as Sora[1], Lumiere[2], CogVideoX[3], KLING[4], MagicVideo[5], and Pika[6], as shown in Fig.Examples of Applications of Video Generation Models.

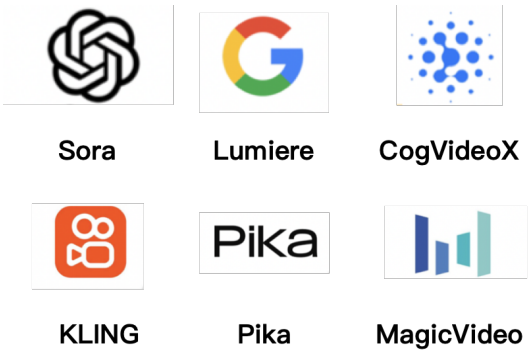

Figure 2: Examples of Applications of Video Generation Models

The generative models behind these products primarily fall into three types:

- **Generative Adversarial Networks (GAN):** The early technical approach to Text-to-Video (T2V) was inspired by Text-to-Image (T2I), using GAN-based models which can be seen as an extension of image generation methods along the temporal dimension. These models consist of two parts: the generator and the discriminator. The goal of the generator is to create videos, while the task of the discriminator is to distinguish which videos are generated fake ones. During the training process, these two networks iteratively improve until the generator is able to produce high-quality videos that are indistinguishable from real ones, and the discriminator can accurately identify the fake videos as much as possible.[7, 8, 9, 10, 11, 12, 13].
- **Autoregressive Transformers:** After the emergence of Transformer, it achieved remarkable results in text generation tasks, leading some researchers to explore its application in video generation tasks. The approach involves using an encoder to convert input modalities, such as text and images, into token sequences. These token sequences represent the information required for generating a video. Once the input modalities are encoded, an autoregressive transformer is then trained to decode each frame of the generated video. This means that the model generates one frame at a time, conditioned on the previous frames and the input tokens, gradually building up the full video sequence. This method allows for efficient learning of temporal dependencies within the video, making it a promising approach for video generation. Research has shown that this technique can yield high-quality results in generating coherent and contextually relevant video sequences, demonstrating the potential of Transformers in the video generation domain [14, 15, 16, 17].

- **Diffusion Models:** In recent years, significant advancements in diffusion models [18, 19]have greatly accelerated the progress of Text-to-Video (T2V) generation. Initially, diffusion-based video generation models were often built upon existing Text-to-Image (T2I) frameworks. These models incorporated a temporal module, allowing them to extend from the image domain to the video domain [20, 21]. This approach enabled the generation of videos by adding time as an additional dimension to the image generation process. However, more recent T2V generation models have evolved to be directly trained on Text-to-Video (T2V) data, bypassing the reliance on T2I models and allowing for more tailored and efficient video generation. These newer models focus on generating coherent video sequences from scratch, learning to capture both the spatial and temporal aspects of video data, and have shown impressive results in generating high-quality videos from textual descriptions. This shift from T2I-based to T2V-focused training has marked a significant step forward in the capabilities of diffusion models in video generation.

These studies provide strong technical support for generating videos consistent with the real world.

2.2 Video Quality Assessment

Video generation incorporates richer physical laws, spatial imaging principles, and dynamic changes over time. As a result, evaluating video generation becomes more complex, requiring consideration of more factors. The research on video evaluation can be categorized as follows:

- **Technical Level:** Traditional Video Quality Assessment (VQA) metrics focus on technical performance aspects, such as compression effects, transmission quality, and playback smoothness[22, 23, 24, 25, 26].
- **Concent Quality Level:** Recent research has shifted its focus to evaluating the concent quality of videos, taking into account factors such as blurriness, motion stability, and noise levels[28, 29].

2.3 Problem Statement

In the productization and commercialization of Text-to-Video(T2V) generative models, the instruction-following capability is a key ability and characteristic, reflecting the consistency between the generated video and its input text.

Therefore, accurately evaluating the instruction-following performance of video generation models is crucial.

However, accurately assessing the instruction-following performance of these models requires extensive manual evaluation. Given the high cost and limited scalability of manual evaluation, it is important to propose an end-to-end reward model to automatically evaluate the performance of text-to-video generation.

3 Method

3.1 Data Pipeline

When constructing datasets for evaluating AI-generated videos, it is important to consider the diversity, quality, and consistency of the data, as well as how to enhance the model's training effectiveness through data integration and transformation. So, we developed the first open-source comprehensive and diverse dataset specifically designed for contrastive learning, named **VideoCross**.

3.1.1 Addressing Data Inconsistency and Redundancy

Below Fig.Several datasets for scoring AI-generated videos listed several significant datasets for scoring AI-generated videos.

A key challenge when building such datasets is the inconsistency that arises from the use of different benchmarks by various research teams for labeling and evaluating models. When constructing video samples, the selection of video generation models should be as diverse as possible to enable the

model to learn content from different data sources. Due to the rapid development in the field of video generation, obtaining high-quality data requires sampling from both state-of-the-art models and earlier models with poorer performance. It's apparent that models like Pika and T2V-Zero have been labeled by different research teams according to their respective benchmarks, and these benchmarks are incompatible during training, resulting in considerable redundancy.

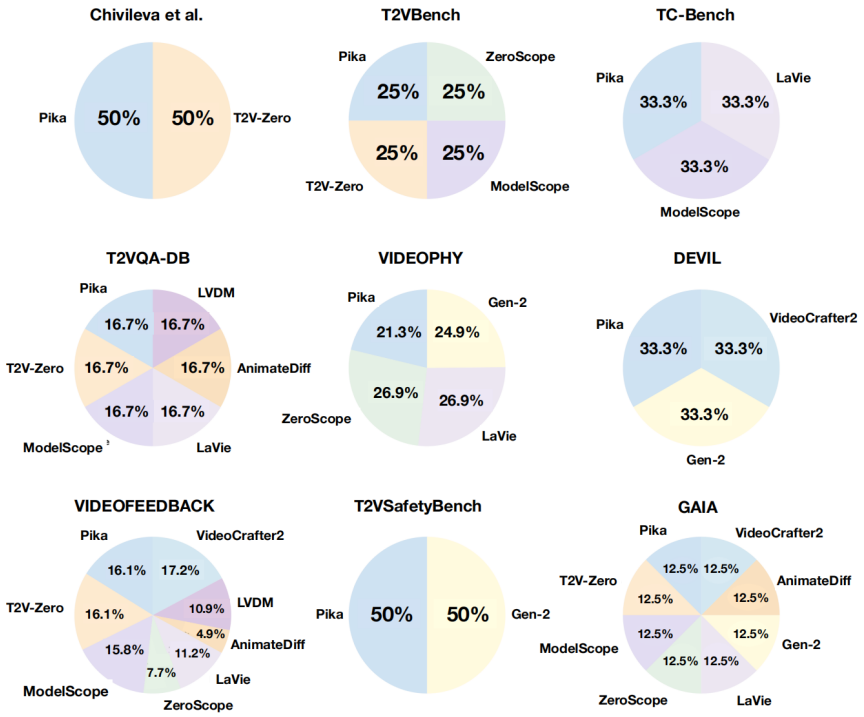

Figure 3: Several datasets for scoring AI-generated videos

To overcome this issue, we introduce **VideoCross**, an open-source dataset specifically designed for contrastive learning. **VideoCross** has $300k$ human preference pairwise videos, integrating dimensions from various benchmarks, merging similar descriptors and eliminating irrelevant ones to reduce redundancy and enhance consistency across the data. This ensures that the dataset is more cohesive and can be used effectively for model training.

3.1.2 Annotation and Sampling Strategy

To further optimize the dataset, the quality of video content is assessed through structured annotation. In **VideoCross**, annotated data is converted into paired data points. Each pair includes a question, two videos, and an answer indicating which video is of higher quality. This setup allows for direct comparison and better evaluation of video generation performance.

In terms of alignment evaluation, we focus on the challenge that early video generation models had in following text instructions, which made it difficult to find two videos with identical text in the dataset. To address this, **VideoCross** adopts a methodology similar to that used by the CLIP[30] model. It constructs numerous positive and negative sample pairs, enabling the model to learn how to generate videos that align with textual descriptions effectively. This approach helps improve the model's understanding of the alignment between text and video.

3.2 Training Method

Contrastive Learning is an unsupervised learning method that aims to learn useful feature representations by maximizing the similarity between samples in the data. The core idea is to bring similar sample pairs (positive pairs) closer together while pushing dissimilar sample pairs (negative pairs) further apart.

In this study, we employ a contrastive learning approach to train the model using positive and negative triplets comprising a better video, a worse video, and corresponding text. This method enhances the model's ability to effectively capture the alignment between videos and text, as illustrated in Fig. Training Method.

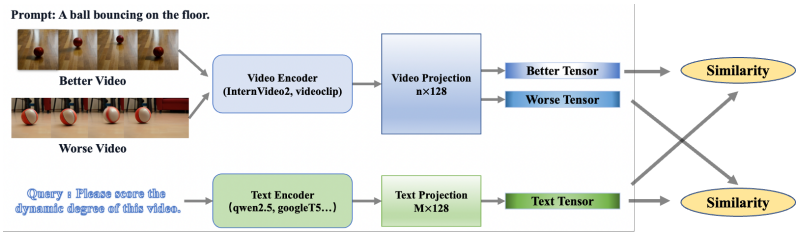

Figure 4: Training Method

3.2.1 Input and Encoding

Video data is fed into the video encoder (e.g., VideoClip or InternVideo2), which processes the input and outputs two video embedding tensors corresponding to the better video and worse video. Text data is input into the text encoder (e.g., Google/T5 or qwen2.5), which outputs a text embedding tensor.

3.2.2 Reward Module

The reward module takes as input two video embedding tensors (corresponding to the better and worse videos) and a single text embedding tensor. The module processes these inputs as follows:

- Each video embedding tensor and the text embedding tensor are independently passed through separate projection layers to map them into a shared 256-dimensional embedding space. This ensures that the representations are comparable and aligned for similarity computation.
- For each video-text pair (better video-text and worse video-text), the cosine similarity is calculated between the corresponding projected embeddings. This operation measures the degree of alignment or compatibility between the video and text representations.
- The output of the reward module consists of two scalar scores—one for the better video-text pair and one for the worse video-text pair. These scores reflect how well each video aligns with the given textual description.

3.2.3 Contrastive Learning and Loss Function

For each input sample pair, contrastive learning is used to train the model. This method aims to maximize the similarity of positive pairs while minimizing the similarity of negative pairs, thereby enabling the model to learn meaningful feature representations. Specifically, for each "video-text" positive and negative sample pair, the loss function is defined as follows:

$$\text{Loss}(\theta) = -E_{(x,y_c,y_r)\sim D} \log\left(\sigma\left(r_\theta(x, y_c) - r_\theta(x, y_r)\right)\right)$$

where $r_\theta(x, y_c)$ represents the score for the "dimension x - video y" pair, where y_c indicates the answer that aligns with human preferences (chosen), and y_r indicates the answer that does not align with human preferences (rejected).

3.2.4 Model Selection and Hardware Configuration

During training, we selected Qwen2-vl-7b-chat and Cogvlm2-video as the base models.

Qwen2-vl and Cogvlm2-video are state-of-the-art models in visual understanding, achieving cutting-edge performance in image and video comprehension tasks.

The training process was conducted using hardware consisting of 16*A800 GPUs to handle the computational demands of large-scale training.

4 Experiment

4.1 Supervised Fine-tuning vs Contrastive Learning

In this section, we explore and compare two common techniques for training our video generation models: Supervised Fine-tuning (SFT) and Contrastive Learning (CL). We provide an overview of both methods, their advantages, and how they are applied in the context of our task.

4.1.1 Supervised Fine-tuning (SFT)

Supervised Fine-tuning involves updating the pre-trained model's weights based on labeled data. In this approach, the model is exposed to a dataset where each video is paired with a corresponding ground truth, allowing the model to learn a direct mapping from input text to video output. We utilize a standard loss function to minimize the discrepancy between predicted and ground truth outputs. This method is particularly useful when large-scale labeled datasets are available and high-accuracy text-to-video alignment is the goal.

4.1.2 Contrastive Learning (CL)

Contrastive Learning, on the other hand, aims to maximize the similarity between positive pairs (e.g., correctly paired text and video) while minimizing the similarity between negative pairs (e.g., mismatched text and video). Unlike supervised learning, CL does not require ground truth video labels, making it more flexible and robust in scenarios with limited labeled data. We employ a contrastive loss function that encourages the model to learn embeddings where semantically related text-video pairs are closer together in the feature space.

The comparison between SFT and CL is critical to understanding their effectiveness in aligning text and video, particularly when working with multimodal data.

4.2 Result Evaluation

In this section, we describe the metrics and methodologies used to evaluate the performance of the models trained with Supervised Fine-tuning (SFT) and Contrastive Learning (CL).

4.2.1 Evaluation Metrics

To assess the effectiveness of the models, we rely on a combination of quantitative and qualitative metrics. The key metrics include:

- **Dynamism**: This metric evaluates the level of dynamic movement in the generated video. It measures how well the generated video captures the intensity and variations in motion based on the textual input, reflecting the vitality of the action.
- **Motion Continuity**: This metric assesses the smoothness and coherence of movement throughout the video. It ensures that actions and transitions between frames are continuous and logically connected, avoiding abrupt or disjointed movements.
- **Visual Aesthetics**: This evaluates the visual quality of the generated video in terms of composition, color grading, framing, and overall artistic appeal. High aesthetic value is critical for videos to be engaging and visually pleasant.
- **Character Continuity**: This metric checks the consistency and continuity of characters throughout the video. It evaluates whether characters remain coherent in appearance, behavior, and positioning across different scenes, without sudden or unexplained changes.
- **Textual Alignment**: This metric measures how well the video aligns with the input textual instructions. It assesses whether the generated video faithfully represents the content, actions, and emotions described in the text, without deviation from the given narrative.
- **Physical Realism**: This evaluates how realistically the generated video adheres to the laws of physics. It examines the plausibility of movements, object interactions, and environmental factors, ensuring that the video does not violate basic physical principles (e.g., gravity, momentum).

These metrics collectively provide a comprehensive assessment of the alignment and quality of the generated videos, ensuring they meet both artistic and technical standards.

Text Input: 3D animated video of john and meeshu travelling in a bicycle inside a deep forest

Output of Reward Model: "visual quality":2,"temporal consistency":2,"dynamic degree":3,"text-to-video alignment":2,"factual consistency":1,

Figure 5: The inference of SFT Model: This figure shows the output of the model after supervised fine-tuning, where the text input is transformed into a video output.

4.2.2 Results Discussion

The results from both training methods (SFT and CL) are compared across the evaluation metrics. Preliminary results suggest that SFT models tend to produce higher fidelity outputs when sufficient labeled data is available, while CL models excel in scenarios with fewer labeled examples. We discuss the trade-offs between these methods in terms of performance, generalization, and data requirements.

4.3 Ablation Study

An ablation study is conducted to evaluate the impact of different components within the model architecture and training process. Specifically, we isolate the effects of:

- **Text Encoding Method**: We compare the use of traditional RNNs versus transformer-based encoders for processing textual input.
- **Video Decoder Architecture**: The impact of different video generation decoders, such as convolutional and GAN-based models, is analyzed.
- **Loss Function Variations**: We experiment with variations of the loss function, including standard MSE loss versus perceptual loss, to gauge their effects on the quality of generated videos.

Through this ablation study, we aim to identify the most effective components of our model architecture and training setup.

5 Conclusion

We introduced CleveReward, a contrastive learning-based reward model. By integrating datasets like T2VQA and videofeedback into a pairwise format and applying contrastive learning, we have demonstrated the model's ability to effectively train across datasets with different scoring systems. While CleveReward is still in development, early results suggest that it has the potential to surpass existing state-of-the-art video reward models.

Our model achieved an accuracy of **69.7%** on the arena-genaibench comparative dataset, which aligns with human preferences and approaches the state-of-the-art (SOTA) algorithms.

This research provides a foundation for automating the evaluation of instruction-following in T2V models, paving the way for more efficient and scalable model optimization. Future work will focus on further refining CleveReward, enhancing its generalizability across diverse datasets, and improving its ability to assess the alignment between generated videos and textual descriptions.

References

[1] OpenAI. *Video Generation Models as World Simulators*. Available at: https://openai.com/index/video-generation-models-as-world-simulators/, 2024. Accessed: October 20, 2024.

[2] O. Bar-Tal, H. Chefer, O. Tov, et al. *Lumiere: A space-time diffusion model for video generation*. arXiv preprint arXiv:2401.12945, 2024.

[3] Z. Yang, J. Teng, W. Zheng, et al. *Cogvideox: Text-to-video diffusion models with an expert transformer*. arXiv preprint arXiv:2408.06072, 2024.

[4] X. Wang, Y. Zhang, Y. Wang, et al. *KLING: Knowledge-driven text-to-video generation for educational content*. arXiv preprint arXiv:2407.13520, 2024.

[5] W. Hong, W. Wang, M. Ding, et al. *MagicVideo: A generative model for high-quality creative video content from text*. arXiv preprint arXiv:2408.16500, 2024.

[6] R. Liu, Y. Fang, F. Yu, et al. *Pika: Real-time interactive text-to-video generation for personalized content*. arXiv preprint arXiv:2408.05718, 2024.

[7] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. *Scaling up GANs for text-to-image synthesis*. Available at: https://arxiv.org/abs/2303.05511, 2023.

[8] T. Karras, S. Laine, and T. Aila. *A style-based generator architecture for generative adversarial networks*. Available at: https://arxiv.org/abs/1812.04948, 2019.

[9] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov. *Generating images from captions with attention*. Available at: https://arxiv.org/abs/1511.02793, 2016.

[10] S. Aigner and M. Körner. *FutureGAN: Anticipating the future frames of video sequences using spatio-temporal 3D convolutions in progressively growing GANs*. Available at: https://arxiv.org/abs/1810.01325, 2018.

[11] P. Bhattacharjee and S. Das. *Temporal coherency based criteria for predicting video frames using deep multi-stage generative adversarial networks*. Advances in Neural Information Processing Systems, 30, 2017.

[12] G. Fox, A. Tewari, M. Elgharib, and C. Theobalt. *StyleVideoGAN: A temporal generative model using a pretrained StyleGAN*. Available at: https://arxiv.org/abs/2107.07224, 2021.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen. *Progressive growing of GANs for improved quality, stability, and variation*. Available at: https://arxiv.org/abs/1710.10196, 2018.

[14] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. *A survey on vision transformer*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1):87–110, 2022.

[15] T. Lin, Y. Wang, X. Liu, and X. Qiu. *A survey of transformers*. Available at: https://arxiv.org/abs/2106.04554, 2021.

[16] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, and N. Duan. *Godiva: Generating open-domain videos from natural descriptions*. Available at: https://arxiv.org/abs/2104.14806, 2021.

[17] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. *Neural discrete representation learning*. Available at: https://arxiv.org/abs/1711.00937, 2018.

[18] J. Ho, A. Jain, and P. Abbeel. *Denoising diffusion probabilistic models*. Advances in Neural Information Processing Systems, 33:6840–6851, 2020.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-resolution image synthesis with latent diffusion models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

[20] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al. *Lavie: High-quality video generation with cascaded latent diffusion models*. arXiv preprint arXiv:2309.15103, 2023.

[21] X. Chen, Y. Wang, L. Zhang, S. Zhuang, X. Ma, J. Yu, Y. Wang, D. Lin, Y. Qiao, and Z. Liu. *Seine: Short-to-long video diffusion model for generative transition and prediction*. arXiv preprint arXiv:2310.20700, 2023.

[22] W. Liu, Z. Duanmu, and Z. Wang. *End-to-end blind quality assessment of compressed videos using deep neural networks*. In Proceedings of the 26th ACM International Conference on Multimedia, MM '18, pages 546–554. ACM, 2018. DOI: https://doi.org/10.1145/3240508.3240643.

[23] L. K. Choi and A. C. Bovik. *Flicker sensitive motion tuned video quality assessment*. In 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), pages 29–32. IEEE, 2016. DOI: https://doi.org/10.1109/SSIAI.2016.7459167.

[24] K. Manasa and S. S. Channappayya. *An optical flow-based full-reference video quality assessment algorithm*. IEEE Transactions on Image Processing, 25(6):2480–2492, 2016. DOI: https://doi.org/10.1109/TIP.2016.2548247.

[25] M. Masry, S. S. Hemami, and Y. Sermadevi. *A scalable wavelet-based video distortion metric and applications*. IEEE Transactions on Circuits and Systems for Video Technology, 16(2):260–273, 2006. DOI: https://doi.org/10.1109/TCSVT.2005.861946.

[26] P. Peng, D. Liao, and Z.-N. Li. *An efficient temporal distortion measure of videos based on spacetime texture*. Pattern Recognition, 70:1–11, 2017. DOI: https://doi.org/10.1016/j.patcog.2017.04.031.

[27] T. Kou, X. Liu, W. Sun, J. Jia, X. Min, G. Zhai, and N. Liu. *StableVQA: A deep no-reference quality assessment model for video stability*. In Proceedings of the 31st ACM International Conference on Multimedia, MM '23, pages 1066–1076. ACM, 2023. DOI: https://doi.org/10.1145/3581783.3611860.

[28] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin. *Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives*. March 2023. Available at: http://arxiv.org/abs/2211.04894. arXiv:2211.04894 [cs, eess].

[29] Y. Wang, S. Inguva, and B. Adsumilli. *YouTube UGC dataset for video compression research*. In IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019.

[30] A. Radford, J. W. Kim, C. Hallacy, et al. *Learning transferable visual models from natural language supervision*. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

[31] J. Xu, X. Liu, Y. Wu, et al. *Imagereward: Learning and evaluating human preferences for text-to-image generation*. Advances in Neural Information Processing Systems, 2024, 36.

[32] E. Fish, J. Weinbren, A. Gilbert. *Two-stream transformer architecture for long video understanding*. arXiv preprint arXiv:2208.01753, 2022.

[33] R. Liu, Y. Fang, F. Yu, et al. *Deep video understanding with video-language model*. In Proceedings of the 31st ACM International Conference on Multimedia, pages 9551–9555, 2023.

[34] X. Wang, Y. Zhang, Y. Wang, et al. *Quo vadis, action recognition? A new model and the Kinetics dataset*. arXiv preprint arXiv:2301.12345, 2023.

[35] W. Hong, W. Wang, M. Ding, et al. *Cogvlm2: Visual language models for image and video understanding*. arXiv preprint arXiv:2408.16500, 2024.

[36] W. Wang, Y. Yang. *VidProm: A million-scale real prompt-gallery dataset for text-to-video diffusion models*. arXiv preprint arXiv:2403.06098, 2024.