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

Anonymous Authors

7 SUPPLEMENTARY

In Sec. 7.1, we provide more details on the establishment of T2VQA-DB, including the design of the subjective study and detailed dataset analysis. In Sec. 7.2, we present more results of T2VQA predictions on T2VQA-DB. The results provide a more comprehensive demonstration of the performance of T2VQA, validating its effectiveness.

7.1 Establishment of T2VQA-DB

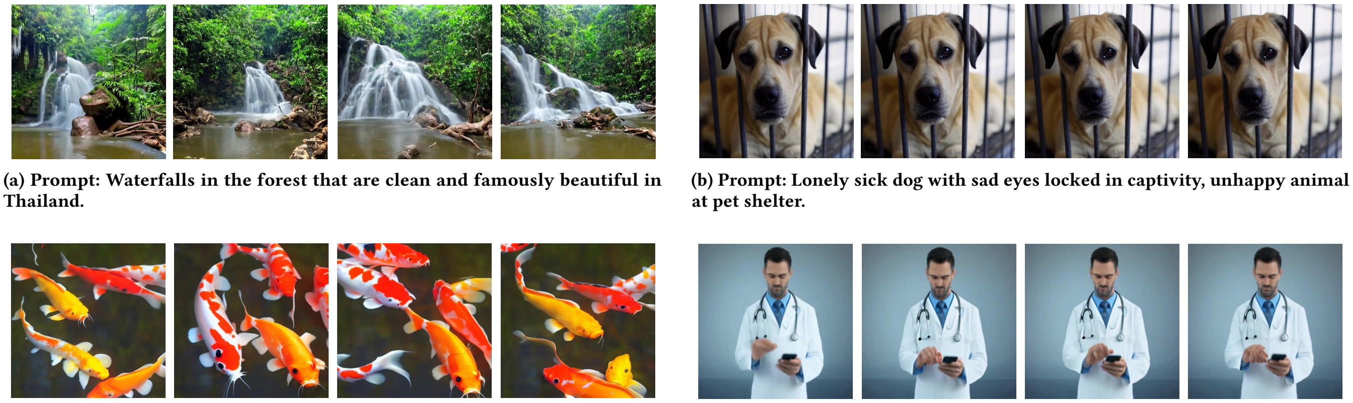
In T2VQA-DB, we collect 10,000 videos generated by 9 different Text-to-Video (T2V) models, including Text2Video-Zero [7], AnimateDiff [3], Tune-a-video [13], VidRD [2] VideoFusion [10], ModelScope [11], LVDM [4], Show-1 [14], and LaVie [12]. We also conduct a subjective study to obtain Mean Opinion Scores (MOS) of each video's overall perceptual quality. Here we introduce the details of the establishment of T2VQA-DB.

7.1.1 Details in Subjective Study. In the subjective study, we invite 27 subjects to score the overall quality of the videos. They are instructed to mainly consider the video quality from two perspectives, in terms of text-video alignment and video fidelity, and then give an overall score to each video. The study is divided into training and testing sessions. The training session helps the subjects to have a general understanding of the scoring standard. We provide 5 example videos with quality from bad to excellent. In the testing session, the subjects score the video using a slider with a range of 0 to 100 and a granularity of 1. We divide the scores into the five ITU-standard [6] grades, *i.e.*bad, poor, fair, good, and excellent, increasing one grade every 20 points. We first show the prompt to let the subjects get familiar with the description, and then the video is played. The subjects give the score after the video is fully played.

After scoring, we conduct quality control to guarantee the reliability of each subject's scores. We split the dataset into 10 groups, with each group 1,000 videos. In each group, we randomly select 50 videos. We calculate Spearman's Rank-Order Correlation Coefficient (SROCC) between each subject's scores and the mean scores of the videos. If the SROCC is lower than 0.6, the subject is required to re-score this group. If the second scoring is still unqualified, the score of this group will be rejected. After the quality control, we consider the remaining scores to be authentic and valid.

7.1.2 Common quality issues. Here we discuss some common factors that lead to video quality degradation. Distortions like noises, blurriness, and high contrast certainly cause quality degradation, as they do in general Video Quality Assessment (VQA) tasks. We do not explain these general distortions in detail here. In the meantime, there are other issues that target text-generated videos. Low consistency between frames is one of them. It refers to the subject or the whole video scene changing greatly between frames, causing a negative effect on the viewer's quality of experience. Fig. 8a shows an example of a video suffering from low consistency. Though each frame of the video matches the prompt's description and has high fidelity, the waterfall and other objects in the scene change in every frame, making the video look chaotic.

Besides, some models may generate videos with irrational parts, which refers to scenes that do not conform to the principles of



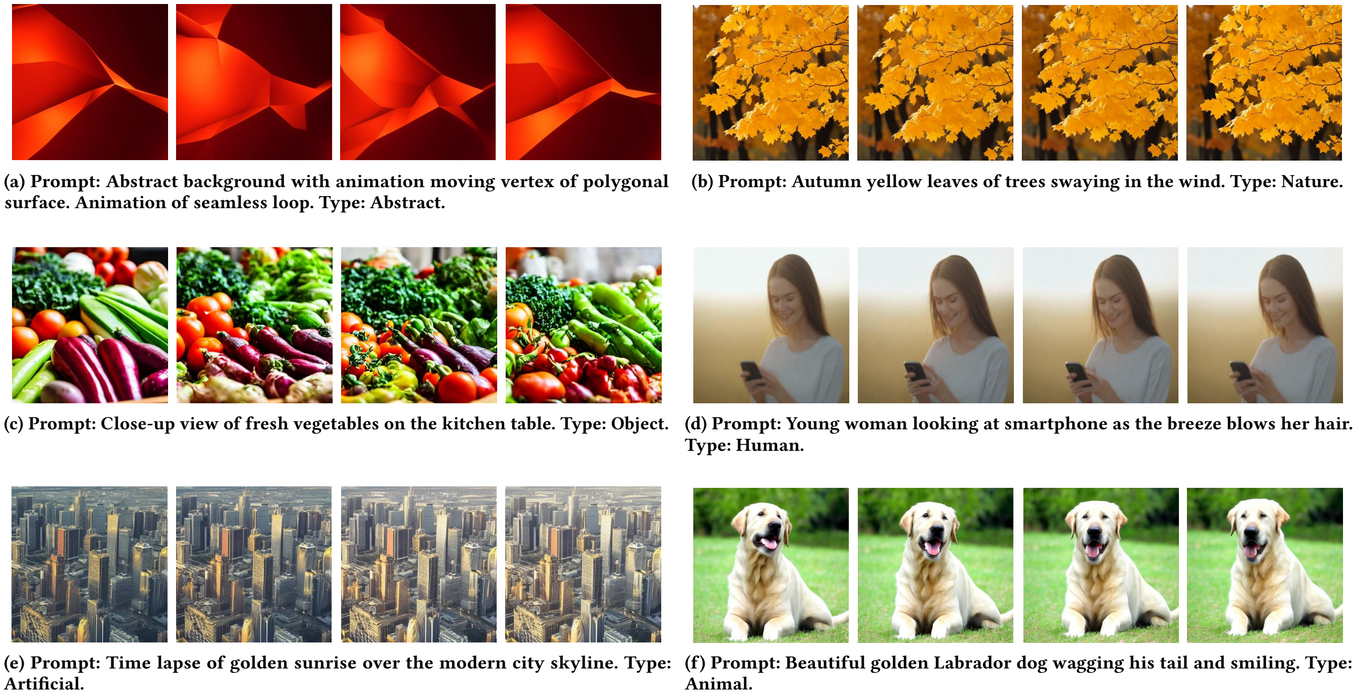
(a) Prompt: Waterfalls in the forest that are clean and famously beautiful in Thailand.

(b) Prompt: Lonely sick dog with sad eyes locked in captivity, unhappy animal at pet shelter.

(c) Prompt: Fish koi carp red, yellow, white, and black float in closed water in the room during the day. Left two girls in the reflection of the water look in the water and take pictures.

(d) Prompt: A male doctor with a mobile phone opens and touches the hologram active ingredient of medicine.

Figure 8: (a): Video example suffers from low consistency. (b): Video example has an irrational problem. (c): Video example has too complicated prompt. (d): Video example has too abstract prompt. The text prompts are listed in the subtitles.



(a) Prompt: Abstract background with animation moving vertex of polygonal surface. Animation of seamless loop. Type: Abstract.

(b) Prompt: Autumn yellow leaves of trees swaying in the wind. Type: Nature.

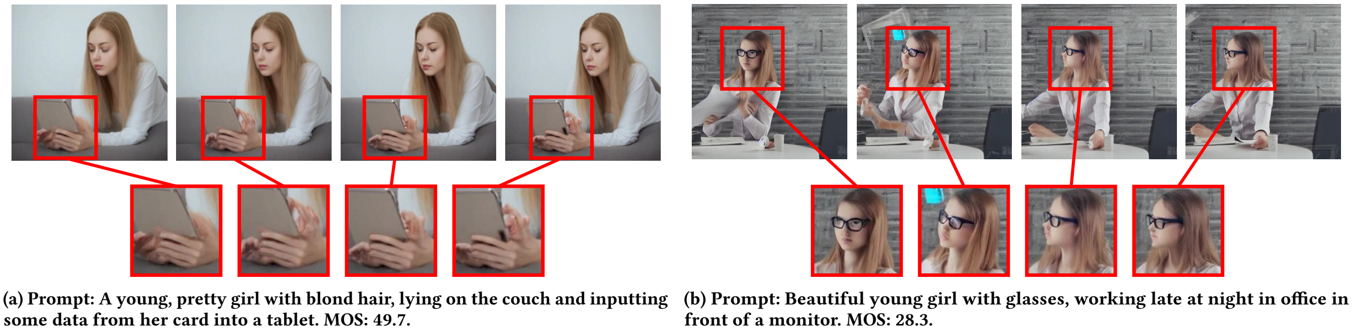
(c) Prompt: Close-up view of fresh vegetables on the kitchen table. Type: Object.

(d) Prompt: Young woman looking at smartphone as the breeze blows her hair. Type: Human.

(e) Prompt: Time lapse of golden sunrise over the modern city skyline. Type: Artificial.

(f) Prompt: Beautiful golden Labrador dog wagging his tail and smiling. Type: Animal.

Figure 9: Video examples from T2VQA-DB. The text prompts and types are listed in the subtitles.



(a) Prompt: A young, pretty girl with blond hair, lying on the couch and inputting some data from her card into a tablet. MOS: 49.7.

(b) Prompt: Beautiful young girl with glasses, working late at night in office in front of a monitor. MOS: 28.3.

Figure 10: Video examples generated by human-type prompts. The text prompts and MOSs are listed in the subtitles.

real physics. Fig 8b shows an example of a dog locked up in a cage. Though the model generates the dog's head and its facial expression well, the dog's head goes through the iron bars, making the video look strange.

Other issues are related to the choice of prompts. If the prompt describes a highly complicated scene, it will be not easy for the models to generate a video strictly matching the description. As shown in Fig. 8c, The prompt describes a scene where two girls taking pictures in the reflection of the water, but the model fails to generate the girls. Another condition is that the prompt describes an unrealistic, or highly fictional scene, which makes the model difficult to understand. The prompt in Fig. 8d requires the model to generate a hologram from the phone, which is a fictional technology. As a result, the model fails to generate the video as required.

7.1.3 Analysis on types. After gaining the MOSs of the videos, we conduct a thorough analysis of the dataset. According to the subject in the prompts' description, we categorize the prompts into nature, artificial, human, animal, object, abstract, and others. Fig. 9 shows video examples from 6 main types. As discussed in the main paper, human-type prompts have the worst performance. For some T2V models, it is difficult for the models to generate the detailed human body structure, such as fingers and facial features. In Fig. 10a, there are multiple fingers generated, while in Fig. 10b, the facial details of the girl are blurry. Both lead to a relatively low score.

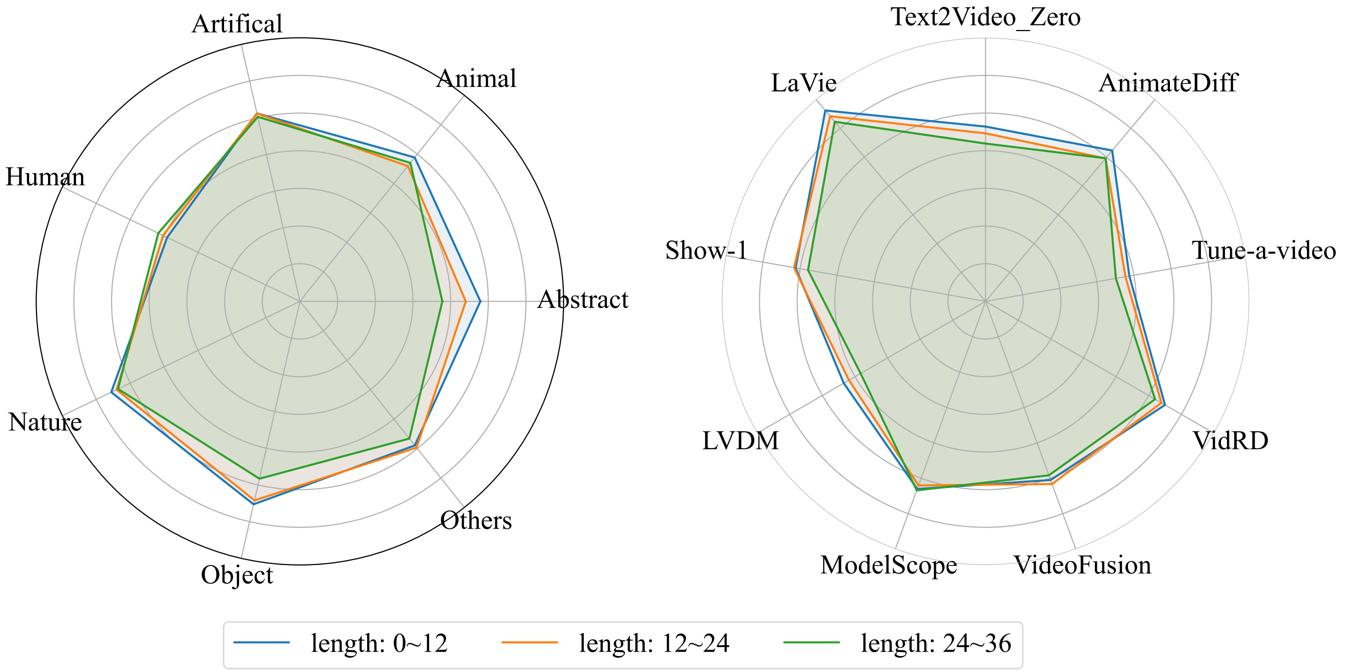


Figure 11: Left: Average scores of different prompt lengths on different prompt types. Right: Models Performance on different prompt lengths.

7.1.4 Analysis on prompt length. We also analyze the lengths of the prompts. The lengths range from 1 to 36 words. We divide the prompts into three groups according to length, with the first group being less than 12 words, the second group being more than 12 and less than 24, and the third more than 24. The partition results in 6130 prompts in the first group, 3060 in the second, and 810 in the third. We analyze the effect of prompt length on different types and models. The results are shown in Fig. 11. Generally, prompts with shorter lengths have a better performance. For different types, the abstract type is the most sensitive to length. The object and others have slightly worse performance on long-length prompts. For different models, the impact of prompt length is not obvious, but there is still a trend of performance decreasing with length.

7.2 More Results of T2VQA

Here we present more results of T2VQA predictions on T2VQA-DB in Fig 12, 13, and 14. We randomly choose 3 prompts and present the 10 videos generated by different T2V models using the same prompt. The MOSs and predictions from T2VQA are listed in the subtitles. The results indicate that T2VQA is able to give relatively accurate predictions, validating its effectiveness.

7.3 Limitations and Future Work

We investigate the limitations in the proposed T2VQA-DB and T2VQA. models like Sora [1] have achieved stunning results with 1080P, 60s high-fidelity videos. Other current T2V models are not capable of generating such high-fidelity videos. We will expand our dataset with Sora generated videos in future work. For T2VQA, since the other existing T2V datasets [5, 8, 9] do not release their subjective scores, we will conduct the cross-dataset validation to validate the generalization of T2VQA over other T2V datasets in future work.

REFERENCES

[1] 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

[2] Jiaxi 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).

[3] 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).

[4] 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).

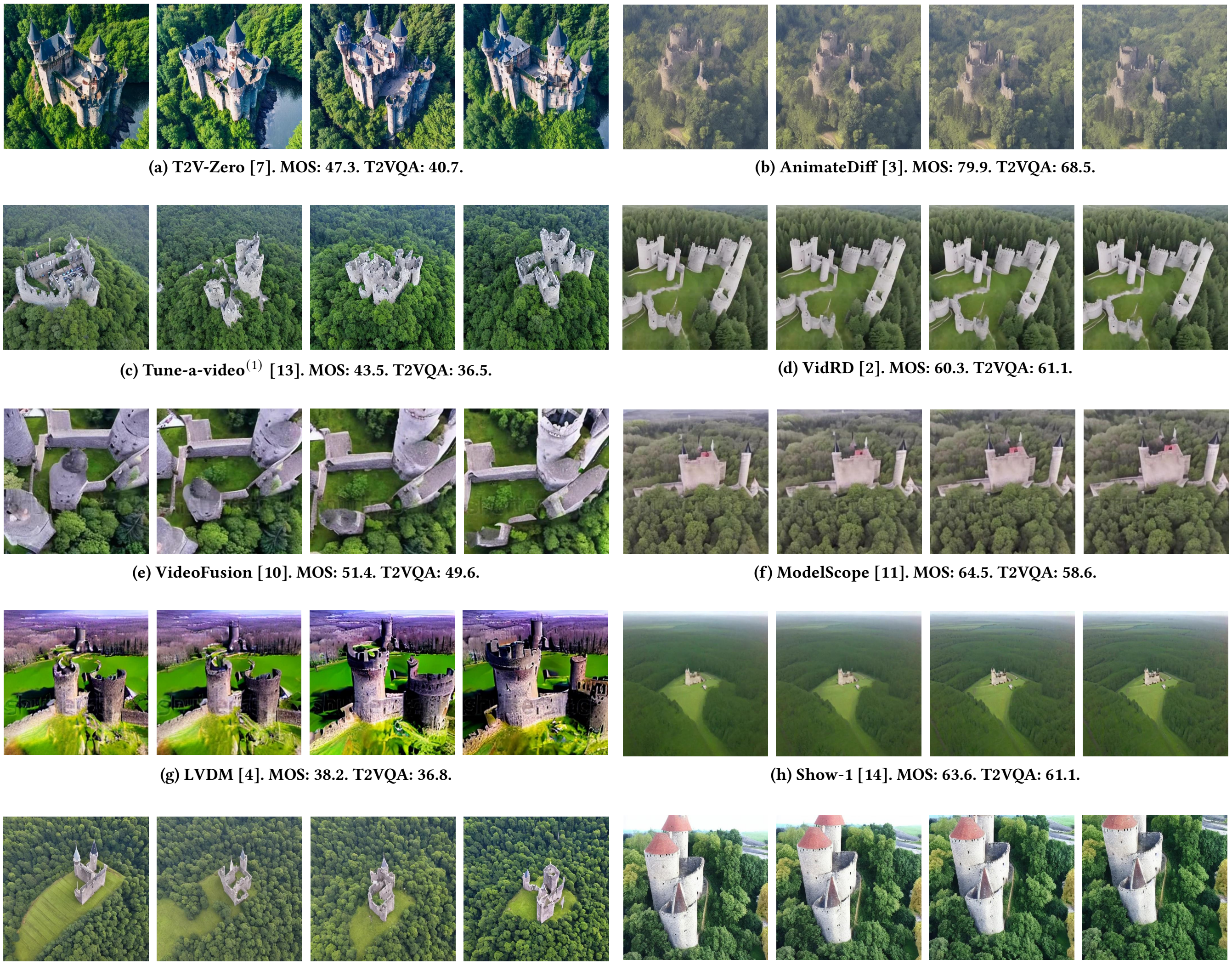
[5] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[6] 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.

[7] 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.

[8] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2023. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440* (2023).

Figure 12: Prompt: Aerial view of an old castle, two towers, and forest around.



(a) T2V-Zero [7]. MOS: 47.3. T2VQA: 40.7.

(b) AnimateDiff [3]. MOS: 79.9. T2VQA: 68.5.

(c) Tune-a-video$^{(1)}$ [13]. MOS: 43.5. T2VQA: 36.5.

(d) VidRD [2]. MOS: 60.3. T2VQA: 61.1.

(e) VideoFusion [10]. MOS: 51.4. T2VQA: 49.6.

(f) ModelScope [11]. MOS: 64.5. T2VQA: 58.6.

(g) LVDM [4]. MOS: 38.2. T2VQA: 36.8.

(h) Show-1 [14]. MOS: 63.6. T2VQA: 61.1.

(i) Tune-a-video$^{(2)}$ [13]. MOS: 40.3. T2VQA: 36.8.

(j) LaVie [12]. MOS: 72.7. T2VQA: 68.2.

[9] 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).

[10] Zhengxiong 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.

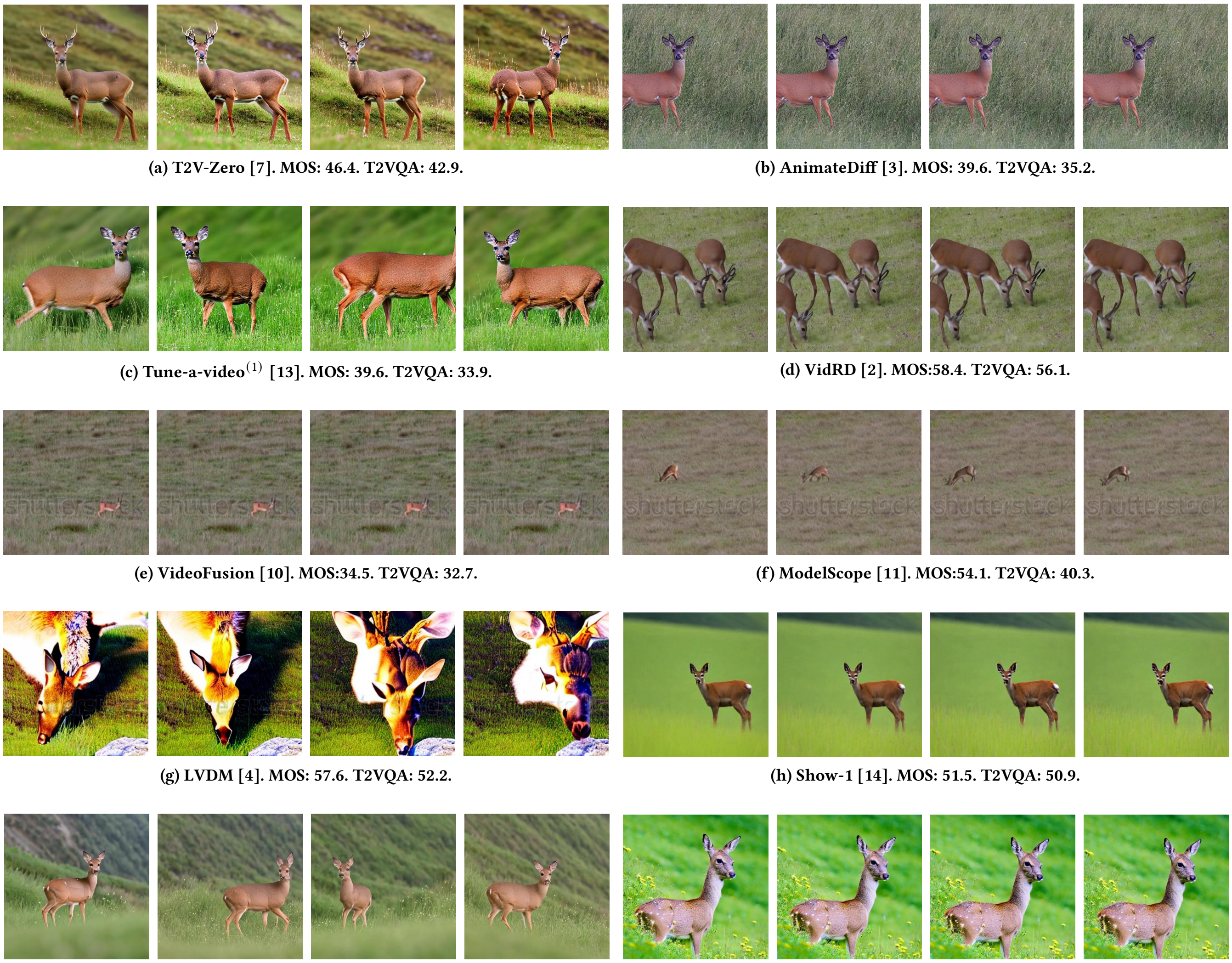
[11] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

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

[13] 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.

[14] 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).

Figure 13: Prompt: Summer in the mountains. mammal species European roe deer (capreolus). a female grazing grass on the mountain meadow.



(a) T2V-Zero [7]. MOS: 46.4. T2VQA: 42.9.

(b) AnimateDiff [3]. MOS: 39.6. T2VQA: 35.2.

(c) Tune-a-video$^{(1)}$ [13]. MOS: 39.6. T2VQA: 33.9.

(d) VidRD [2]. MOS:58.4. T2VQA: 56.1.

(e) VideoFusion [10]. MOS:34.5. T2VQA: 32.7.

(f) ModelScope [11]. MOS:54.1. T2VQA: 40.3.

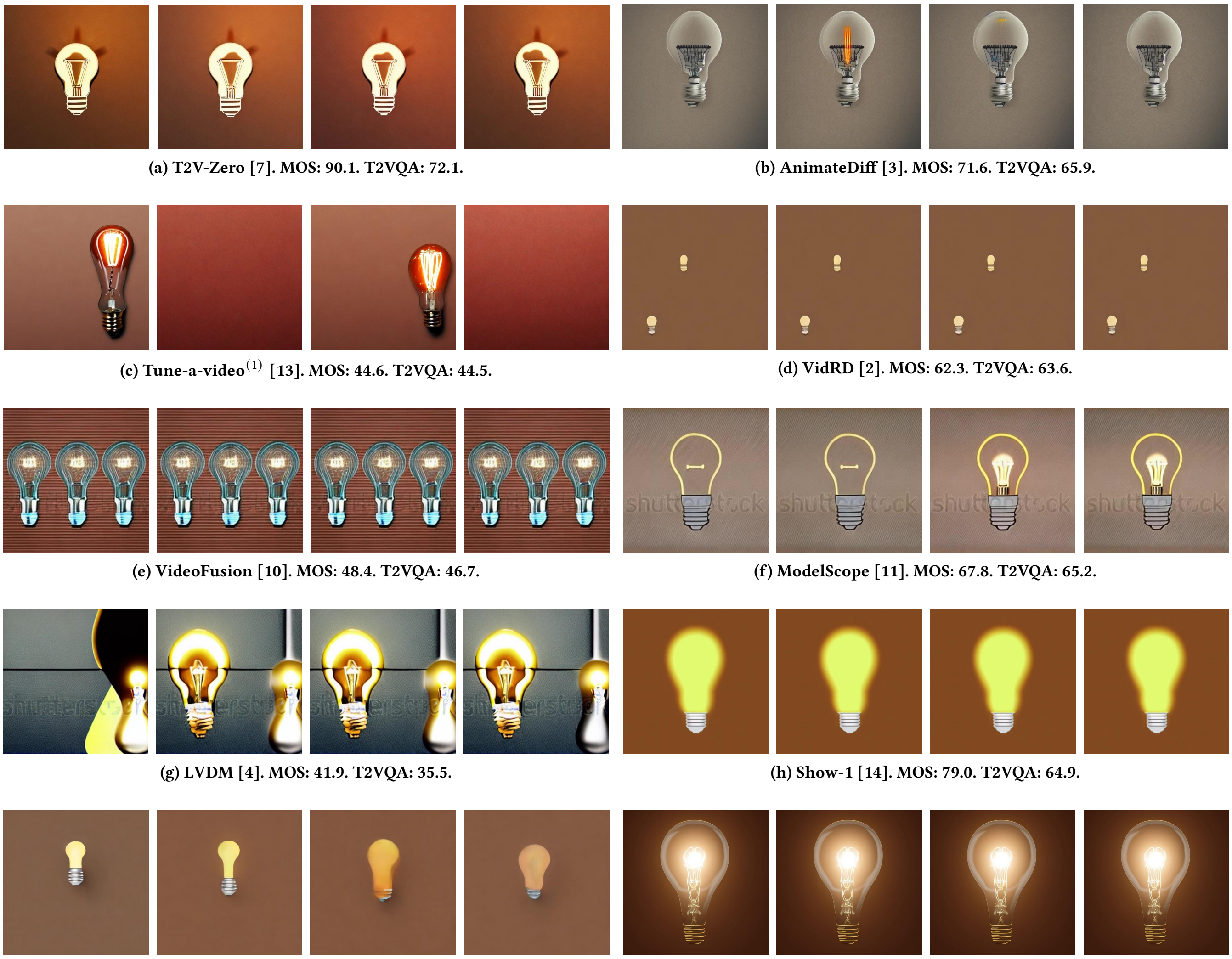
(g) LVDM [4]. MOS: 57.6. T2VQA: 52.2.

(h) Show-1 [14]. MOS: 51.5. T2VQA: 50.9.

(i) Tune-a-video$^{(2)}$ [13]. MOS: 37.9. T2VQA: 35.5.

(j) LaVie [12]. MOS: 78.2. T2VQA: 64.6.

Figure 14: Prompt: Light bulb animation on brown background.



(a) T2V-Zero [7]. MOS: 90.1. T2VQA: 72.1.

(b) AnimateDiff [3]. MOS: 71.6. T2VQA: 65.9.

(c) Tune-a-video$^{(1)}$ [13]. MOS: 44.6. T2VQA: 44.5.

(d) VidRD [2]. MOS: 62.3. T2VQA: 63.6.

(e) VideoFusion [10]. MOS: 48.4. T2VQA: 46.7.

(f) ModelScope [11]. MOS: 67.8. T2VQA: 65.2.

(g) LVDM [4]. MOS: 41.9. T2VQA: 35.5.

(h) Show-1 [14]. MOS: 79.0. T2VQA: 64.9.

(i) Tune-a-video$^{(2)}$ [13]. MOS: 58.6. T2VQA: 56.0.

(j) LaVie [12]. MOS: 82.7. T2VQA: 76.8.