

EvalCrafter: Benchmarking and Evaluating Large Video Generation Models

Yaofang Liu^{1,2,*} Xiaodong Cun^{1,*} Xuebo Liu³ Xintao Wang¹ Yong Zhang¹ Haoxin Chen¹
 Yang Liu⁴ Tieyong Zeng⁴ Raymond Chan^{2†} Ying Shan¹

¹ Tencent AI Lab ² City University of Hong Kong

³ University of Macau ⁴ The Chinese University of Hong Kong

<http://evalcrafter.github.io>

Abstract

The vision and language generative models have been overgrown in recent years. For video generation, various open-sourced models and public-available services are released for generating high-visual quality videos. However, these methods often use a few academic metrics, e.g., FVD [56] or IS [46], to evaluate the performance. We argue that it is hard to judge the large conditional generative models from the simple metrics since these models are often trained on very large datasets with multi-aspect abilities. Thus, we propose a new framework and pipeline to exhaustively evaluate the performance of the generated videos. To achieve this, we first conduct a new prompt list for text-to-video generation by analyzing the real-world prompt list with the help of the large language model. Then, we evaluate the state-of-the-art video generative models on our carefully designed benchmarks, in terms of visual qualities, content qualities, motion qualities, and text-caption alignment with around 18 objective metrics. To obtain the final leaderboard of the models, we also fit a series of coefficients to align the objective metrics to the users' opinions. Based on the proposed opinion alignment method, our final score shows a higher correlation than simply averaging the metrics, showing the effectiveness of the proposed evaluation method.

1. Introduction

The charm of the large generative models is sweeping the world. *e.g.*, the well-known ChatGPT and GPT4 [39] have shown human-level abilities in several aspects, including coding, solving math problems, and even visual understanding, which can be used to interact with our human beings using any knowledge in a conversational way. As for the



Figure 1. We propose EvalCrafter, a comprehensive framework for benchmarking and evaluating the text-to-video models, including the well-defined prompt types in grey and the multiple evaluation aspects in black circles.

generative models for visual content creation, Stable Diffusion [44] and SDXL [41] play very important roles since they are the most powerful publicly available models that can generate high-quality images from any text prompts.

Beyond text-to-image, taming diffusion model for video generation has also progressed rapidly. Early works (Imagen-Viedo [22], Make-A-Video [49]) utilize the cascaded models for video generation directly. Powered by the image generation priors in Stable Diffusion, LVDM [20] and MagicVideo [69] have been proposed to train the temporal layers to efficiently generate videos. Apart from the academic papers, several commercial services also can generate videos from text or images. *e.g.*, Gen2 [16] and PikaLabs [4]. Although we can not get the technique details of these services, they are not evaluated and compared with other methods.

*Equal contribution.

†Corresponding author.

However, all current large text-to-video (T2V) model only uses previous GAN-based metrics, like FVD [56], for evaluation, which only concerns the distribution matching between the generated video and the real videos, other than the pairs between the text prompt and the generated video. Differently, we argue that a good evaluation method should consider the metrics in different aspects, *e.g.*, the motion quality and the temporal consistency. Also, similar to the large language models, some models are not publicly available and we can only get access to the generated videos, which further increases the difficulties in evaluation. Although the evaluation has progressed rapidly in the large generative models, including the areas of LLM [39], MLLM [33], and text-to-image [25], it is still hard to directly use these methods for video generation. The main problem here is that different from text-to-image or dialogue evaluation, motion and consistency are very important to video generation which previous works ignore.

We make the very first step to evaluate the large multi-modality generative models for video. In detail, we first build a comprehensive prompt list containing various everyday objects, attributes, and motions. To achieve a balanced distribution of well-known concepts, we start from the well-defined meta-types of the real-world knowledge and utilize the knowledge of large language models, *e.g.*, ChatGPT [39], to extend our meta-prompt to a wide range. Besides the prompts generated by the model, we also select the prompts from real-world users and text-to-image prompts. After that, we also obtain the metadata (*e.g.*, color, size, *etc.*) from the prompt for further evaluation usage. Second, we evaluate the performance of these larger T2V models from different aspects, including the video visual qualities, the text-video alignment, and the motion quality and temporal consistency. For each aspect, we use one or more objective metrics as the evaluation metrics. Since these metrics only reflect one of the abilities of the model, we also conduct a multi-aspects user study to judge the model in terms of its qualities. After obtaining these opinions, we train the coefficients of each objective regression model to align the evaluation scores to the user's choice, so that we can obtain the final scores of the models and also evaluate the new video using the trained coefficients.

Overall, we summarize the contribution of our paper as:

- We make the first step of evaluating the large T2V model and build a comprehensive prompt list with detailed annotations for T2V evaluation.
- We consider the aspects of the video visual quality, video motion quality, and text-video alignment for the evaluation of video generation. For each aspect, we align the opinions of humans and also verify the effectiveness of the proposed metric by human alignment.

- During the evaluation, we also discuss several conclusions and findings, which might be also useful for further training of the T2V generation models.

2. Related Work

2.1. Text-to-Video Generation and Evaluation

T2V generation aims to generate videos from the given text prompts. Early works generate the videos through Variational AutoEncoders (VAEs [28]) or generative adversarial network (GAN [18]). However, the quality of the generated videos is often low quality or can only work on a specific domain, *e.g.*, face [66] or landscape [50, 63]. With the rapid development of the diffusion model [23], video diffusion model [24], and large-scale text-image pretraining [43], current methods utilize the stronger text-to-image pre-trained model prior to generation. *e.g.*, Make-A-Video [49] and Imagen-Video [22] train a cascaded video diffusion model to generate the video in several steps. LVDM [20], Align Your latent [8] and MagicVideo [69] extend the latent text-to-image model to video domains by adding additional temporal attention or transformer layer. AnimateDiff [19] shows a good visual quality by utilizing the personalized text-to-image model. Similar methods are also been proposed by SHOW-1 [65] and LAVIE [59]. T2V generation also raises the enthusiasm of commerce or non-commerce companies. For online model services, *e.g.*, Gen1 [16] and Gen2 [16], show the abilities of the high-quality generated video in the fully T2V generation or the conditional video generation. For discord-based servers, Pika-Lab [4], Morph Studio [3], FullJourney [2] and Floor33 Pictures [20] also show very competitive results. Besides, there are also some popular open-sourced text (or image)-to-video models, *e.g.*, Zero-Scope [5], ModelScope [57].

However, these methods still lack a fair and detailed benchmark to evaluate the advantages of each method. For example, they only evaluate the performance using FVD [56] (LVDM [20], MagicVideo [69], Align Your Latent [8]), IS [46] (Align Your Latent [8]), CLIP similarity [43] (Gen1 [16], Imagen Video [22], Make-A-Video [49]), or user studies to show the performance level. These metrics might only perform well on previous in-domain text-to-image generation methods but ignore the alignment of input text, the motion quality, and the temporal consistency, which are also important for T2V generation.

2.2. Evaluations on Large Generative Models

Evaluating the large generative models [39, 41, 44, 54, 55] is a big challenge for both the NLP and vision tasks. For the large language models, current methods design several metrics in terms of different abilities, question types, and user platform [13, 21, 61, 68, 70]. More details of LLM evaluation and Multi-model LLM evaluation can be found in

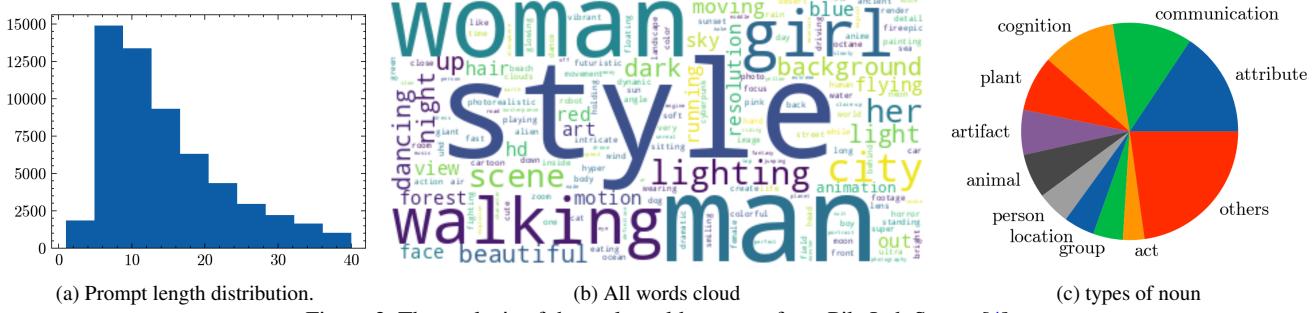


Figure 2. The analysis of the real-world prompts from PikaLab Server [4].

recent surveys [10, 67]. Similarly, the evaluation of the multi-modal generative model also draws the attention of the researchers [7, 62]. For example, Seed-Bench [33] generates the VQA for multi-modal large language model evaluation.

For the models in visual generation tasks, Imagen [45] only evaluates the model via user studies. DALL-Eval [12] assesses the visual reasoning skills and social basis of the text-to-image model via both user and object detection algorithm [9]. HRS-Bench [6] proposes a holistic and reliable benchmark by generating the prompt with ChatGPT [39] and utilizing 17 metrics to evaluate the 13 skills of the text-to-image model. TIFA [25] proposes a benchmark utilizing the visual question answering (VQA). However, these methods still work for text-to-image evaluation or language model evaluation. For T2V evaluation, we consider the quality of motion and temporal consistency.

3. Benchmark Construction

Our benchmark aims to create a trustworthy prompt list to evaluate the abilities of various of T2V models fairly. To achieve this goal, we first collect and analyze the T2V prompt from large-scale real-world users. After that, we propose an automatic pipeline to increase the diversity of the generated prompts so that they can be identified and evaluated by pre-trained computer vision models. Since video generation is time-consuming, we collect 500 prompts as our initial version for evaluation with careful annotation. Below, we give the details of each step.

3.1. What Kind of Prompts Should We Generate?

To answer this question, we collect the prompts from the real-world T2V generation discord users, including the FullJourney [2] and PikaLab [4]. In total, we get over 600k prompts with corresponding videos and filter them to 200k by removing repeated and meaningless prompts. Our first curiosity is how long a prompt should be generated, as shown in Fig. 2 (a), 90% of the prompts contain the words in the range of [3, 40]. We also plot the most important words in Fig. 2 (b) by removing some unclear words like video, camera, high, quality, etc., where the person, the style, human motion, and scene are dominant. Despite the above analysis,

we also count the word class to decide the meta class of our prompt list. As shown in Fig. 2 (c), we use WordNet [37] to identify the meta classes, except for the communication, attribute, and cognition words, the artifacts (human-made objects), human, animal, and the location (landscape) play important roles. We also add the most important word style of Fig. 2 (b) to the metaclass. Overall, we divide the T2V generation into roughly four meta-subject classes, including the human, animal, object, and landscape. For each type, we also consider the motions and styles of each type and the relationship between the current metaclass and other metaclasses to construct the video. Besides, we include the motion which is relevant to the main object and important for the video. Finally, we consider the camera motion and the style by template.

3.2. General Recognizable Prompt Generation

Automatically Prompt Generation. After deciding the meta classes of our prompt list, we generate the recognizable prompt by the power of a large language model (LLM) and humans. As shown in Fig 3, for each kind of meta class, we let GPT-4 [39] describe the scenes about this meta class with randomly sampled meta information along with the attributes of the scenes so that we already know the labels. For example, for humans, we can ask GPT-4 to give us the attributes of humankind, age, gender, clothes, and human activity, which are saved as a JSON file as the ground truth computer vision models. However, we also find that the GPT-4 is not fully perfect for this task, the generated attributes are not very consistent with the generated description. Thus, we involve a self-check to the benchmark building, where we also use GPT-4 to identify the similarities of the generated description and the meta data. Finally, we filter the prompts by ourselves to make sure each prompt is correct and meaningful for T2V generation.

Prompts from Real World. Since we have already collected a very large scale of prompts from real-world users and there are also available text-to-image evaluation prompts, *e.g.*, DALL-Eval [12] and Draw-Bench [45], we also integrate these prompts to our benchmark list. To achieve this, we first filter and generate the metadata using GPT-4. Then, we

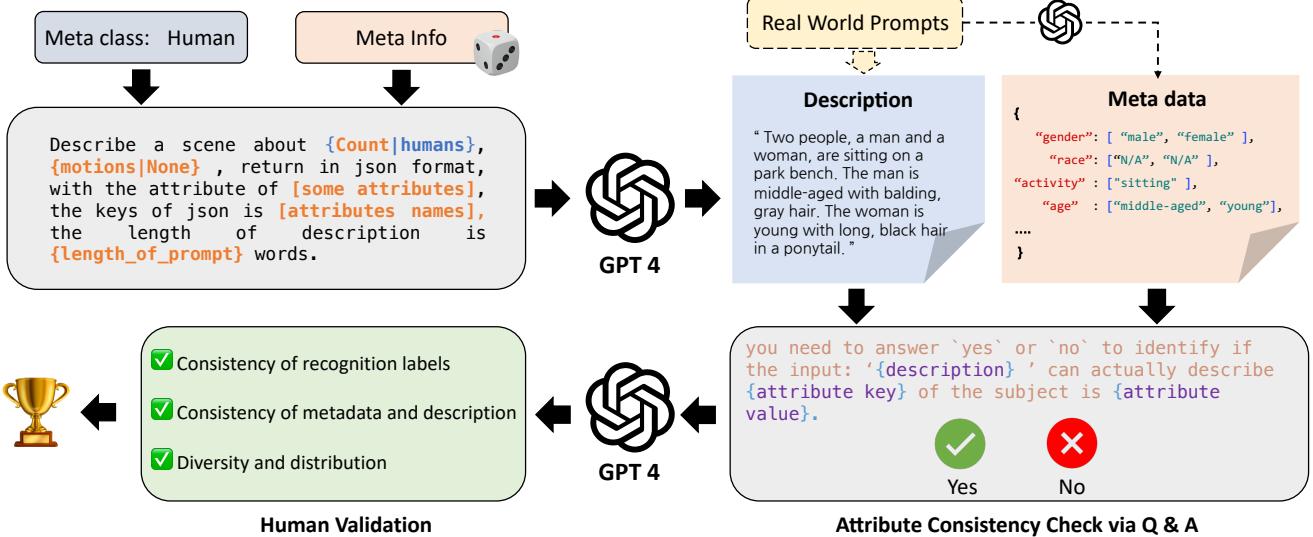


Figure 3. We aim to generate a trustworthy benchmark with detailed prompts for text-to-video evaluation by computer vision model and users. We show the pipeline above.

choose the suitable prompts with the corresponding meta-information as shown in Fig. 3 and check the consistency of the meta-information.

3.3. Benchmark Analysis

Overall, we get over 500 prompts in the meta classes of human, animal, objects, and landscape. Each class contains the natural scenes, the stylized prompts, and the results with explicit camera motion controls. We give a brief view of the benchmark in Fig. 4. The whole benchmark contains over 500 prompts with careful categories. To increase the diversity of the prompts, our benchmark contains 3 different sub-types as shown in Figure 4, where we have a total of 50 styles and 20 camera motion prompts. We add them randomly in the 50% prompts of the whole benchmark. Our benchmark contains an average length of 12.5 words pre-prompt, which is also similar to the real-world prompts as we find in Figure. 2.

4. Evaluation Metrics

Different from previous FID [47] based evaluation metrics, we evaluate the T2V models in different aspects, including the visual quality of the generated video, the text-video alignment, the content correctness, the motion quality, and temporal consistency. Below, we give the detailed metrics.

4.1. Overall Video Quality Assessment

We first consider the visual quality of the generated video, which is the key for visually appealing to the users. Notice that, since the distribution-based method, *e.g.*, FVD [56] still needs the ground truth video for evaluation, we argue

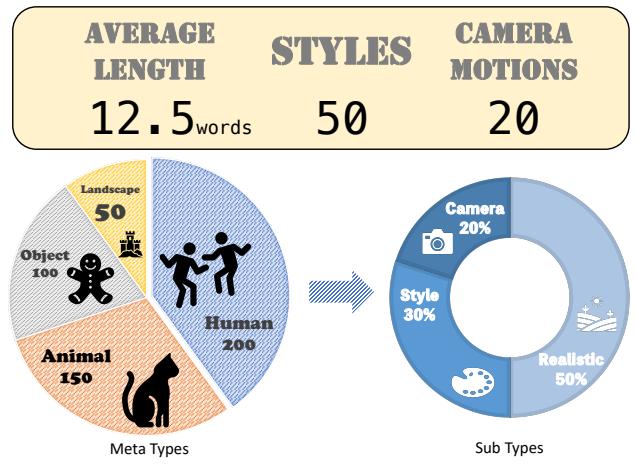


Figure 4. The analysis of the proposed benchmarks. Each meta type contains three sub-types to increase the diversity of the generated videos.

these kinds of metrics are not suitable for the general T2V generation cases.

Video Quality Assessment (VQA_A , VQA_T). We utilize the state-of-the-art video quality assessment method, Dover [60], to evaluate the quality of the generated video in terms of aesthetics and technicality, where the technical rating measures the quality of the generated video in terms of the common distortion, including noises, artifacts, *etc.* Dover [60] is trained on a self-collected larger-scale dataset and the labels are ranked by the real users for alignment. We term the aesthetic and technical scores as VQA_A and VQA_T , respectively.

Inception Score (IS). Following previous metrics in the T2V

Method	Ver.	Abilities [†]	Resolution	FPS	Open Source	Length	Speed*	Motion	Camera
ModelScope	23.03	T2V	256×256	8	✓	4s	0.5 min	-	-
VideoCrafter	23.04	T2V	256×256	8	✓	2s	0.5 min	-	-
ZeroScope	23.06	T2V & V2V	1024×576	8	✓	4s	3 min	-	-
ModelScope-XL	23.08	I2V & V2V	1280×720	8	✓	4s	8 min+	-	-
Floor33 Pictures	23.08	T2V	1280×720	8	-	2s	4 min	-	-
PikaLab	23.09	I2V OR T2V	1088×640	24	-	3s	1 min	✓	✓
Gen2	23.09	I2V OR T2V	896×512	24	-	4s	1 min	✓	✓

Table 1. The difference in the available diffusion-based text-to-video models. [†] We majorly evaluate the method of text-to-video generation (T2V). For related image-to-video generation model (I2V), *i.e.*, ModelScope-XL, we first generate the image by Stable Diffusion v2.1 and then perform image-to-video on the generated content.

generation papers, we also use the inception score [46] of the video as one of the video quality assessment indexes. The inception score is proposed to evaluate the performance of GAN [18], which utilizes a pre-trained Inception Network [52] on the ImageNet [15] dataset as the pre-trained feature extraction method. The inception score reflects the diversity of the generated video, whereas a larger score means the generated content is more diverse.

4.2. Text-Video Alignment

Another common evaluation direction is the alignment of the input text and the generated video. We not only consider both the global text prompts and the video, and also the content correctness in different aspects. Below, we give the details of each score.

Text-Video Consistency (CLIP-Score). We incorporate the CLIP-Score as one of the evaluation metrics, given its widespread usage and simplicity in quantifying the discrepancy between input text prompts and generated videos. Utilizing the pretrained ViT-B/32 CLIP model [43] as a feature extractor, we obtain frame-wise image embeddings and text embeddings, and compute their cosine similarity. The cosine similarity for the t -th frame of the i -th video x_t^i and the corresponding prompt p^i is denoted as $\mathcal{C}(\text{emb}(x_t^i), \text{emb}(p^i))$, $\text{emb}(\cdot)$ means CLIP embedding. The overall CLIP-Score, S_{CS} , is derived by averaging individual scores across all frames and videos, calculated as

$$S_{CS} = \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{N} \sum_{t=1}^N \mathcal{C}(\text{emb}(x_t^i), \text{emb}(p^i)) \right), \quad (1)$$

where M is the total number of testing videos and N is the total number of frames in each video.

Image-Video Consistency (SD-Score). Most current video diffusion models are fine-tuned on a base stable diffusion with a larger scale dataset. Also, tuning the new parameters for stable diffusion will cause conceptual forgetting, we thus propose a new metric by comparing the generated quality with the frame-wise stable diffusion [44]. In detail, we use SDXL [41] to generate N_1 images $\{d_k\}_{k=1}^{N_1}$ for every prompt

and extract the visual embeddings in both generated images and video frames, and here we set N_1 to 5. We calculate the embedding similarity between the generated videos and the SDXL images, which is helpful to ablate the concept forgotten problems when fine-tuning the text-to-image diffusion model to video models. The final SD-Score is

$$S_{SD} = \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{N} \sum_{t=1}^N \left(\frac{1}{N_1} \sum_{k=1}^{N_1} \mathcal{C}(\text{emb}(x_t^i), \text{emb}(d_k^i)) \right) \right). \quad (2)$$

Text-Text Consistency (BLIP-BLEU). We also consider the evaluation between the text descriptions of the generated video and the input text prompt. To this purpose, we utilize BLIP2 [35] for caption generation. Similar to text-to-image evaluation methods [6], we use BLEU [40] for text alignment of the generated and the source prompt across frames:

$$S_{BB} = \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{N_2} \sum_{k=1}^{N_2} \mathcal{B}(p^i, l_k^i) \right), \quad (3)$$

where $\mathcal{B}(\cdot, \cdot)$ is the BLEU similarity scoring function, $\{l_k^i\}_{k=1}^{N_2}$ are BLIP generated captions for i -th video, and N_2 is set to 5 experimentally.

Object and Attributes Consistency (Detection-Score, Count-Score and Color-Score). For general objects, we employ a state-of-the-art segmentation and tracking method, namely SAM-Track [11], to analyze the correctness of the video content that we are interested in. Leveraging the powerful segmentation model [29], we can easily obtain the objects and their attributes. In our pipeline, we focus on detecting prompts with COCO classes [36], which is a widely used dataset for object detection and segmentation tasks. We evaluate T2V models on the existence of objects, as well as the correctness of color and count of objects in text prompts. Specifically, we assess the Detection-Score, Count-Score, and Color-Score as follows:

1. **Detection-Score (S_{Det})**: Measures average object pres-

ence across videos, calculated as:

$$S_{Det} = \frac{1}{M_1} \sum_{i=1}^{M_1} \left(\frac{1}{N} \sum_{t=1}^N \sigma_t^i \right), \quad (4)$$

where M_1 is the number of prompts with objects, and σ_t^i is the detection result for frame t in video i (1 if an object is detected, 0 otherwise).

2. *Count-Score* (S_{Count}): Evaluates average object count difference, calculated as:

$$S_{Count} = \frac{1}{M_2} \sum_{i=1}^{M_2} \left(1 - \frac{1}{N} \sum_{t=1}^N \frac{|c_t^i - \hat{c}^i|}{\hat{c}^i} \right), \quad (5)$$

where M_2 is the number of prompts with object counts, c_t^i is the detected object count frame t in video i and \hat{c}^i is the ground truth object count for video i .

3. *Color-Score* (S_{Color}): Assesses average color accuracy, calculated as:

$$S_{Color} = \frac{1}{M_3} \sum_{i=1}^{M_3} \left(\frac{1}{N} \sum_{t=1}^N s_t^i \right), \quad (6)$$

where M_3 is the number of prompts with object colors, s_t^i is the color accuracy result for frame i in video t (1 if the detected color matches the ground truth color, 0 otherwise).

Human Analysis (Celebrity ID Score). Human is important for the generated videos as shown in our collected real-world prompts. To this end, we also evaluate the correctness of human faces using DeepFace [48], a popular face analysis toolbox. We do the analysis by calculating the distance between the generated celebrities' faces with corresponding real images of the celebrities.

$$S_{CIS} = \frac{1}{M_4} \sum_{i=1}^{M_4} \left(\frac{1}{N} \sum_{t=1}^N \left(\min_{k \in \{1, \dots, N_3\}} \mathcal{D}(x_t^i, f_k^i) \right) \right), \quad (7)$$

where M_4 is the number of prompts that contain celebrities, $\mathcal{D}(\cdot, \cdot)$ is the Deepface's distance function, $\{f_k^i\}_{k=1}^{N_3}$ are collected celebrities images for prompt i , and N_3 is set to 3.

Text Recognition (OCR-Score) Another hard case for visual generation is to generate the text in the description. To examine the abilities of current models for text generation, we utilize the algorithms from Optical Character Recognition (OCR) models similar to previous text-to-image evaluation [6] or multi-model LLM evaluation method [33]. Specifically, we utilize PaddleOCR¹ to detect the English text generated by each model. Then, we calculate Word Error Rate (WER) [30], Normalized Edit Distance (NED) [51], Character Error Rate (CER) [38], and finally we average these three score to get the OCR-Score.

¹<https://github.com/PaddlePaddle/PaddleOCR>

4.3. Motion Quality

For video, we believe the motion quality is a major difference from other domains, such as image. To this end, we consider the quality of motion as one of the main evaluation metrics in our evaluation system. Here, we consider two different motion qualities introduced below.

Action Recognition (Action-Score). For videos about humans, we can easily recognize the common actions via pre-trained models. In our experiments, we use MMAction2 toolbox [14], specifically the pre-trained VideoMAE V2 [58] model, to infer the human actions in the generated videos. We then take the classification accuracy (ground truth are actions in the input prompts) as our Action-Score. In this work, we focus on Kinetics 400 action classes [26], which is widely used and encompasses human-object interactions like playing instruments and human-human interactions, including handshakes and hugs.

Average Flow (Flow-Score). We also consider the general motion information of the video. To this end, we use the pretrained optical flow estimation method, RAFT [53], to extract the dense flows of the video in every two frames. Then, we calculate the average flow on these frames to obtain the average flow score of every specific generated video clip since some methods are likely to generate still videos which are hard to identify by the temporal consistency metrics.

Amplitude Classification Score (Motion AC-Score). Based on the average flow, we further identify whether the motion amplitude in the generated video is consistent with the amplitude specified by the text prompt. To this end, we set an average flow threshold ρ that if surpasses ρ , one video will be considered large, and here ρ is set to 2 based on our subjective observation. We mark this score to identify the movement of the generated video.

4.4. Temporal Consistency

Temporal consistency is also a very valuable field in our generated video. To this end, we involve several metrics for calculation. We list them below.

Warping Error. We first consider the warping error, which is widely used in previous blind temporal consistency methods [31, 32, 42]. In detail, we first obtain the optical flow of each two frames using the pre-trained optical flow estimation network [53], then, we calculate the pixel-wise differences between the warped image and the predicted image. We calculate the warp differences on every two frames and calculate the final score using the average of all the pairs.

Semantic Consistency (CLIP-Temp). Besides pixel-wise error, we also consider the semantic consistency between every two frames, which is also used in previous video editing works [16, 42]. Specifically, we consider the semantic embeddings on each of the two frames of the generated videos and then get the averages on each two frames, which is shown as

Dimensions	Metrics	ModelScope-XL [57]	ZeroScope [5]	Floor33 [20]	PikaLab [4]	Gen2 [16]
Video Quality	VQA _A \uparrow	97.72	95.95	98.11	99.32	99.04
	VQA _T \uparrow	6.09	6.50	7.60	<u>8.69</u>	10.13
	IS \uparrow	15.99	13.35	<u>15.10</u>	13.66	12.57
Text-video Alignment	CLIP-Score \uparrow	20.62	20.20	21.15	20.72	20.90
	BLIP-BLUE \uparrow	<u>22.42</u>	21.20	23.67	21.89	22.33
	SD-Score \uparrow	68.50	67.79	69.04	<u>69.14</u>	69.31
	Detection-Score \uparrow	49.59	45.80	55.00	50.49	52.44
	Color-Score \uparrow	<u>40.10</u>	46.35	35.07	36.57	32.29
	Count-Score \uparrow	47.67	47.88	57.63	56.46	57.19
	OCR Score \downarrow	83.74	82.58	81.09	<u>81.33</u>	92.94
	Celebrity ID Score \uparrow	<u>45.66</u>	45.96	45.24	43.43	44.58
Motion Quality	Action Score \uparrow	<u>73.75</u>	71.74	74.48	69.84	54.99
	Motion AC-Score \rightarrow	26.67	53.33	60.00	40.00	40.00
	Flow-Score \rightarrow	2.28	1.66	2.23	0.11	0.18
Temporal Consistency	CLIP-Temp \uparrow	99.72	99.84	99.58	99.97	99.92
	Warping Error \downarrow	73.04	80.32	69.77	<u>66.88</u>	58.19
	Face Consistency \uparrow	98.89	<u>99.33</u>	99.05	99.64	99.06

Table 2. Raw results from the aspects of video quality, text-video alignment, motion quality, and temporal consistency.

follows:

$$S_{CT} = \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{N-1} \sum_{t=1}^{N-1} \mathcal{C}(\text{emb}(x_t^i), \text{emb}(x_{t+1}^i)) \right), \quad (8)$$

Face Consistency. Similar to CLIP-Temp, we evaluate the human identity consistency of the generated videos. Specifically, we select the first frame as the reference and calculate the cosine similarity of the reference frame embedding with other frames' embeddings. Then, we average the similarities as the final score:

$$S_{FC} = \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{N-1} \sum_{t=1}^{N-1} \mathcal{C}(\text{emb}(x_{t+1}^i), \text{emb}(x_1^i)) \right), \quad (9)$$

4.5. User Opinion Alignments

Besides the above objective metrics, we conduct user studies on the main five aspects to get the users' opinions. These aspects include (1) *Video Qualities*. It indicates the quality of the generated video where a higher score shows there is no blur, noise, or other visual degradation. (2) *Text and Video Alignment*. This opinion considers the relationships between the generated video and the input text-prompt, where a generated video has the wrong count, attribute, and relationship will be considered as low-quality samples. (3) *Motion Quality*. In this metric, the users need to identify the correctness of the generated motions from the video. (4) *Temporal Consistency*. Temporal consistency is different from motion quality. In motion quality, the user needs to give a rank for high-quality movement. However, in temporal

consistency, they only need to consider the frame-wise consistency of each video. (5) *Subjective likeness*. This metric is similar to the aesthetic index, a higher value indicates the generated video generally achieves human preference, and we leave this metric used directly.

For evaluation, we generate videos using the provided prompts benchmark on five state-of-the-art methods of ModelScope [57], ZeroScope [5], Gen2 [16], Floor33 [1], and PikaLab [4], getting 2.5k videos in total. For a fair comparison, we change the aspect ratio of Gen2 and PikaLab to 16 : 9 to suitable other methods. Also, since PikaLab can not generate the content without the visual watermark, we add the watermark of PikaLab to all other methods for a fair comparison. We also consider that some users might not understand the prompt well, for this purpose, we use SDXL [41] to generate three reference images of each prompt to help the users understand better, which also inspires us to design an SD-Score to evaluate the models' text-video alignments. For each metric, we ask three users to give opinions between 1 to 5, where a large value indicates better alignments. We use the average score as the final labeling and normalize it to range [0, 1].

Upon collecting user data, we proceed to perform human alignment for our evaluation metrics, with the goal of establishing a more reliable and robust assessment of T2V algorithms. Initially, we conduct alignment on the data using the mentioned individual metrics above to approximate human scores for the user's opinion in the specific aspects. We employ a linear regression model to fit the parameters in each dimension, inspired by the works of the evaluation

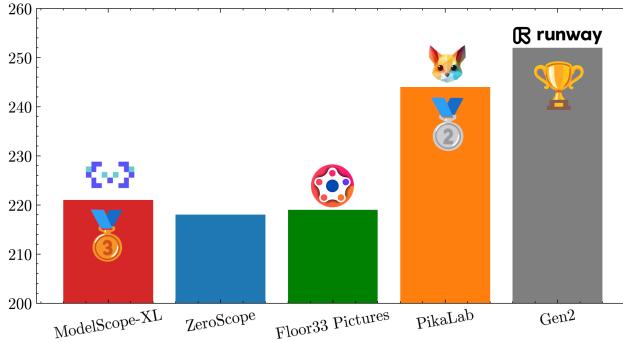


Figure 5. Overall comparison results on our EvalCrafter benchmark.

	Visual Quality	Text-Video Alignment	Motion Quality	Temporal Consistency
ModelScope-XL	55.23 (5)	47.22 (4)	59.41 (2)	59.31 (4)
ZeroScope	56.37 (4)	46.18 (5)	54.26 (4)	61.19 (3)
Floor33 Pictures	59.53 (3)	51.29 (3)	51.97 (5)	56.36 (5)
PikaLab	63.52 (2)	54.11 (1)	57.74 (3)	69.35 (2)
Gen2	67.35 (1)	52.30 (2)	62.53 (1)	69.71 (1)

Table 3. Human-preference aligned results from four different aspects, with the rank of each aspect in the brackets.

of natural language processing [17, 34]. Specifically, we randomly choice 300 samples from four different methods as the fittings samples and left the rest 200 samples to verify the effectiveness of the proposed method (as in Table. 4). The coefficient parameters are obtained by minimizing the residual sum of squares between the human labels and the prediction from the linear regression model. In the subsequent stage, we integrate the aligned results of these four aspects and calculate the average score to obtain a comprehensive final score, which effectively represents the performance of the T2V algorithms. This approach streamlines the evaluation process and provides a clear indication of model performance.

5. Results

We conduct the evaluation on 500 prompts from our benchmark prompts, where each prompt has a metafile for additional information as the answer of evaluation. We generate the videos using all available high-resolution T2V models, including the ModelScope [57], Floor33 Pictures [1], and ZeroScope [5]. We keep all the hyper-parameters, such as classifier-free guidance, as the default value. For the service-based model, we evaluate the performance of the representative works of Gen2 [16] and PikaLab [4]. They generate at least 512p videos with high-quality watermark-free videos. Before our evaluation, we show the differences between each video type in Table 1, including the abilities of these models, the generated resolutions, and fps. As for the comparison on speed, we run all the available models on an NVIDIA A100. For the unavailable model, we run their model online and measure the approximate time. Notice that, PikaLab [4] and Gen2 [16] also have the ability to control the motions and

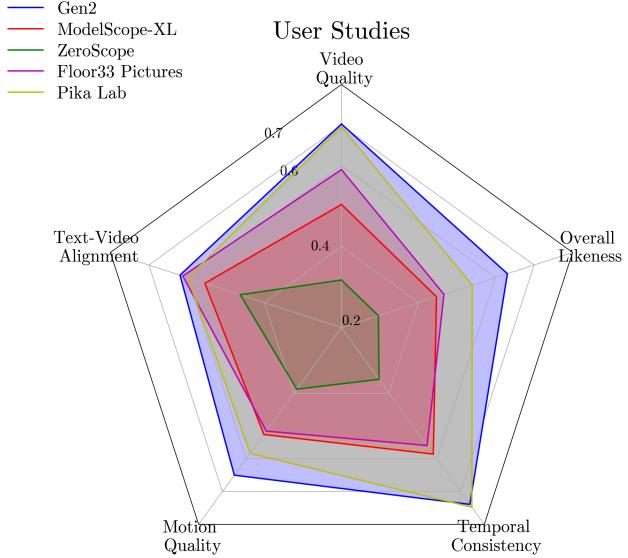


Figure 6. The raw ratings from our user studies.

the cameras through additional hyper-parameters. Besides, although there are many parameters that can be adjusted, we keep the default settings for a relatively fair comparison.

We first show the overall human-aligned results in Fig. 5, with also the different aspects of our benchmark in Table 3, which gives us the final and the main metrics of our benchmark. Finally, as in Figure 7, we give the results of each method on four different meta-types (*i.e.*, animal, human, landscape, object) in our benchmark and two different type videos (*i.e.*, general, style) in our benchmark. For comparing the objective and subjective metrics of each method, we give the raw data of each metric in Table. 1 and Fig. 6. We give a detailed analysis in Sec. 5.1.

5.1. Analysis

Finding #1: Evaluating the model using one single metric is unfair. From Table. 3, the rankings of the models vary significantly across these aspects, highlighting the importance of a multi-aspect evaluation approach for a comprehensive understanding of their performance. For instance, while Gen2 outperforms other models in terms of Visual Quality, Motion Quality, and Temporal Consistency, PikaLab demonstrates superior performance in Text-Video Alignment.

Finding #2: Evaluating the models’ abilities by meta-type is necessary. As shown in Fig. 7, most methods show very different values in different meta types. For example, although Gen2 [16] has the best overall T2V alignment in our experiments, the generated videos from this method are hard to recognize by the action recognition models. We subjectively find Gen2 [16] mainly generates the close-up shot from text prompt with a weaker motion amplitude.

Finding #3: Users are more tolerant with the bad T2V alignment than visual quality. As shown in Fig. 7 and

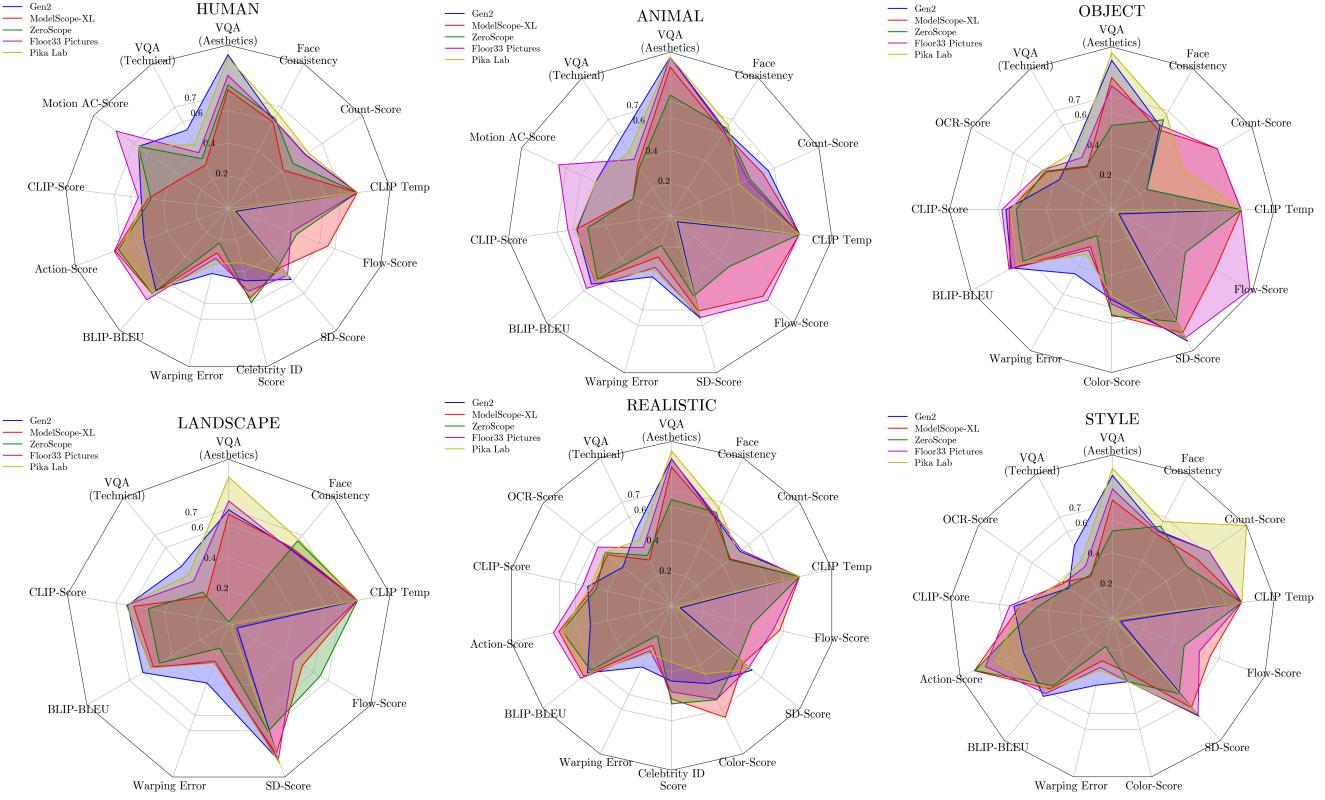


Figure 7. **Raw results in different aspects.** We consider 4 main meta types (animal, human, landscape, object) to evaluate the performance of the meta types of the generated video, where each type contains several prompts with fine-grained attribute labels. For each prompt, we also consider the style of the video, yet more diverse prompts, as shown in **realistic** and **style** figure above. (The metrics values are normalized for better visualization, we preprocess the warping Error and OCR-score, so for these two metrics, a large value indicates better performance in this figure.)

Table. 2, even Gen2 [16] can not perform well in all the text-video alignment metrics, the user still likes the results of this model in most cases due to its good temporal consistency, visual quality, and small motion amplitude.

Finding #4: All the methods CAN NOT control their camera motion directly from the text prompt. Although some additional hyper-parameters can be set as additional control handles, the text encoder of the current T2V text encoder still lacks the understanding of the reasoning behind open-world prompts, like camera motion.

Finding #5: Visually appealing has no positive correlation with the generated resolutions. As shown in Tab. 1, gen2 [16] has the smallest resolutions, however, both humans and the objective metrics consider this method to have the best visual qualities and few artifacts as in Tab. 2, Fig. 6.

Finding #6: Larger motion amplitude does not indicate a better model for users. From Fig. 6, both two small motion models, *i.e.*, PikaLab [4] and Gen2 [16] get better scores in the user’s choice than the larger motion model, *i.e.*, Floor33 Pictures [1]. Where users are more likely to see slight movement videos other than a video with bad and

unreasonable motions.

Finding #7: Generating text from text descriptions is still hard. Although we report the OCR-Scores of these models, we find it is still too hard to generate realistic fonts from the text prompts, nearly all the methods are fair to generate high-quality and consistent texts from text prompts.

Finding #8: The current video generation model still generates the results in a single shot. All methods show a very high consistency of CLIP-Temp as in Table. 2, which means each frame has a very similar semantic across frames. So the current T2V models are more likely to generate the cinemagraphs, other than the long video with multiple transitions and actions.

Finding #9: Most valuable objective metrics. By aligning the objective metrics to the real users, we also find some valuable metrics from a single aspect. For example, SD-Score and CLIP-Score are both valuable for text-video alignment according to Table. 2 and Table. 3. VQA_T and VQA_A are also valuable for visual quality assessment.

Finding #10: Gen2 is not perfect also. Although Gen2 [16] achieved the overall top performance in our evaluation,

Aspects	Methods	Spearsman's ρ	Kendall's ϕ
Visual Quality	VQA _A	42.1	30.5
	VQA _T	49.3	35.9
	Avg.	45.9	33.7
	Ours	50.2	37.6
Motion Amplitude	Motion AC	-16.9	-13.1
	Flow-Score	-32.9	-23.1
	Avg.	-27.8	-20.4
	Ours	32.1	24.0
Temporal Consistency	CLIP-Temp.	50.0	35.8
	Warp Error	36.1	27.1
	Avg.	37.2	27.9
	Ours	50.0	36.0
TV Alignment	SD-Score	10.0	6.9
	CLIP-Score	14.4	10.1
	Avg.	20.2	14.0
	Ours	30.5	21.7

Table 4. **Correction Analysis.** Correlations between some objective metrics and human judgment on text-to-video generations. We use Spearman's ρ and Kendall's ϕ for correlation calculation.

it still has multiple problems. For example, Gen2 [16] is hard to generate video with complex scenes from prompts. Gen2 [16] has a weird identity for both humans and animals, which is also reflected by the IS metric (hard to be identified by the network also) in Table. 1, while other methods do not have such problems.

Finding #11: A significant performance gap exists between open-source and closed-source T2V models. Referring to Table 3, we can observe that open-source models such as ModelScope-XL and ZeroScope have lower scores in almost every aspect compared to closed-source models like PikaLab [4] and Gen2 [16]. This indicates that there is still room for improvement in open-source T2V models to reach the performance levels of their closed-source counterparts.

5.2. Ablation on Human Preference Alignment

To demonstrate the effectiveness of our model in aligning with human scores, we calculate Spearman's rank correlation coefficient [64] and Kendall's rank correlation coefficient [27], both of which are non-parametric measures of rank correlation. These coefficients provide insights into the strength and direction of the association between our method results and human scores, as listed in Table. 4. From this table, the proposed weighting method shows a better correlation on the unseen 200 samples than directly averaging (we divide all data by 100 to get them to range [0, 1] first). Another interesting finding is that all current Motion Amplitude scores are not related to the users' choice. We argue that humans care more about the stability of the motion than

the amplitude. However, our fitting method shows a higher correlation.

5.3. Limitation

Although we have already made a step in evaluating the T2V generation, there are still many challenges. (i) Currently, we only collect 500 prompts as the benchmark, where the real-world situation is very complicated. More prompts will show a more detailed benchmark. (ii) Evaluating the motion quality of the general senses is also hard. However, in the era of multi-model LLM and large video foundational models, we believe better and larger video understanding models will be released and we can use them as our metrics. (iii) The labels used for alignment are collected from only 3 human annotators, which may introduce some bias in the results. To address this limitation, we plan to expand the pool of annotators and collect more diverse scores to ensure a more accurate and unbiased evaluation.

6. Conclusion

Discovering more abilities of the open world large generative models is essential for better model design and usage. In this paper, we make the very first step for the evaluation of the large and high-quality T2V models. To achieve this goal, we first built a detailed prompt benchmark for T2V evaluation. On the other hand, we give several objective evaluation metrics to evaluate the performance of the T2V models in terms of the video quality, the text-video alignment, the object, and the motion quality. Finally, we conduct the user study and propose a new alignment method to match the user score and the objective metrics, where we can get final scores for our evaluation. The experiments show the abilities of the proposed methods can successfully align the users' opinions, giving the accurate evaluation metrics for the T2V methods.

References

- [1] Floor33 pictures discord server. <https://www.morphstudio.com/>. Accessed: 2023-08-30. 7, 8, 9
- [2] Fulljourney discord server. <https://www.fulljourney.ai/>. Accessed: 2023-08-30. 2, 3
- [3] Morph studio discord server. <https://www.morphstudio.com/>. Accessed: 2023-08-30. 2
- [4] Pika Lab discord server. <https://www.pika.art/>. Accessed: 2023-08-30. 1, 2, 3, 7, 8, 9, 10
- [5] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed: 2023-08-30. 2, 7, 8
- [6] EslamMohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan-Farooq Khan, LiErran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. Apr 2023. 3, 5, 6
- [7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 3
- [8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [10] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023. 3
- [11] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 5
- [12] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. 3
- [13] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *arXiv preprint arXiv:2305.14938*, 2023. 2
- [14] MMAAction2 Contributors. Openmmlab’s next generation video understanding toolbox and benchmark. <https://github.com/open-mmlab/mmaction2>, 2020. 6
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5
- [16] Patrick Esser, Johnathan Chiu, Parmida Atighchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1, 2, 6, 7, 8, 9, 10
- [17] Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, and David Traum. Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 726–734, 2020. 8
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 5
- [19] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2
- [20] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. 1, 2, 7
- [21] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 2
- [22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2
- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2
- [25] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 2, 3
- [26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6
- [27] Maurice George Kendall. Rank correlation methods. 1948. 10
- [28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5
- [30] Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002. 6
- [31] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 6

- [32] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020. 6
- [33] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. Jul 2023. 2, 3, 6
- [34] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129:1238–1257, 2021. 8
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5
- [36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5
- [37] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3
- [38] Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*, 2004. 6
- [39] OpenAI. Gpt-4 technical report, 2023. 1, 2, 3
- [40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 5
- [41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 5, 7
- [42] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023. 6
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 5
- [45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3
- [46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1, 2, 5
- [47] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, August 2020. Version 0.3.0. 4
- [48] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended light-face: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 6
- [49] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 2
- [50] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhosiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 2
- [51] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 6
- [52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 5
- [53] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2
- [56] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1, 2, 4
- [57] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 7, 8
- [58] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking, 2023. 6

- [59] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiahuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. 2
- [60] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *International Conference on Computer Vision (ICCV)*, 2023. 4
- [61] Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. On the tool manipulation capability of open-source large language models, 2023. 2
- [62] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3
- [63] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. 2
- [64] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7, 2005. 10
- [65] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. 2
- [66] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, 2022. 2
- [67] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 3
- [68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2
- [69] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1, 2
- [70] Gu Zhouhong, Zhu Xiaoxuan, Ye Haoning, Zhang Lin, Wang Jianchen, Jiang Sihang, Xiong Zhuozhi, Li Zihan, He Qianyu, Xu Rui, Huang Wenhao, Zheng Weiguo, Feng Hongwei, and Xiao Yanghua. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. *arXiv:2304.11679*, 2023. 2