EvalCrafter: Benchmarking and Evaluating Large Video Generation Models

Yaofang Liu[1,2,*] Xiaodong Cun[1,*] Xuebo Liu[3] Xintao Wang[1]
Yong Zhang[1] Haoxin Chen[1] Yang Liu[4†] Tieyong Zeng[4] Raymond Chan[2†] Ying Shan[1]
[1] Tencent AI Lab [2] City University of Hong Kong [3] University of Macau
[4] The Chinese University of Hong Kong
Project Page: 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 have been developed to generate high-quality videos. However, these methods often use a few metrics, e.g.*, FVD [56] or IS [45], 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 novel framework and pipeline for exhaustively evaluating the performance of the generated videos. Our approach involves generating a diverse and comprehensive list of 700 prompts for text-to-video generation, which is based on an analysis of real-world user data and generated with the assistance of a large language model. Then, we evaluate the state-of-the-art video generative models on our carefully designed benchmark, in terms of visual qualities, content qualities, motion qualities, and text-video alignment with 17 well-selected objective metrics. To obtain the final leaderboard of the models, we further fit a series of coefficients to align the objective metrics to the users' opinions. Based on the proposed human 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 [37] 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 generative models for visual content creation, Stable Diffusion (SD) [43] and SDXL [40] play very important roles since they are the most powerful publicly available models that can generate high-quality images from any text prompts.

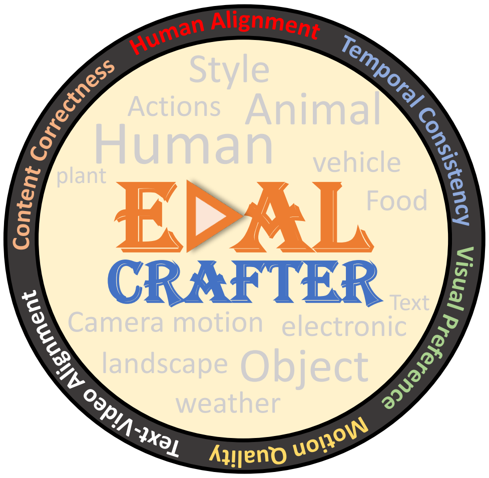


Figure 1. We propose EvalCrafter, a comprehensive framework for benchmarking and evaluating the text-to-video models, including the well-defined prompt types in grey and the multiple evaluation aspects in black circles.

Beyond text-to-image (T2I), taming diffusion model for video generation has also progressed rapidly. Early works (Imagen-Viedo [23], Make-A-Video [49]) utilize the cascaded models for video generation directly. Powered by the image generation priors in SD, LVDM [21] 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 [18] and PikaLabs [5]. Although we can not get the technique details of these services, they are not evaluated and compared with other methods. However, all current large text-to-video (T2V) models only use 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

[*] Equal Contribution.
[†] Corresponding Author

metrics in different aspects, *e.g*., the motion quality and the temporal consistency. Also, similar to the large language models (LLMs), 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 [37], MLLM [33], and T2I [26], it is still hard to directly use these methods for video generation. The main problem here is that different from T2I 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 general T2V models. 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 the LLM, *i.e*., ChatGPT [37], 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 T2I prompts. After that, we obtain the metadata (*e.g*., color, size, *etc*.) from the prompt for further evaluation. Second, we assess the performance of large T2V models from four aspects, *i.e*., video quality, text-video alignment, motion quality, and temporal consistency. For each aspect, we employ several objective metrics as evaluation measures, and we conduct a user study to human scores w.r.t. these four aspects. After that, we train coefficients of the regression model for each aspect, aligning evaluation scores with user preferences. This enables us to obtain the final model scores and evaluate new videos 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, video temporal consistency, 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 correlation analysis.
- During the evaluation, we also discuss several conclusions and findings, which might also contribute to further innovation and development of T2V models.

2. Related Work

2.1. Text-to-Video Generation and Evaluation

Text-to-video (T2V) generation aims to create videos from text prompts. Early works used Variational AutoEncoders (VAEs [29]) or generative adversarial networks (GANs [20]) but often yielded low-quality or domain-specific results, such as faces [66] or landscapes [50, 63]. Recent methods leverage advancements in diffusion models [24, 25, 60] and large-scale text-image pretraining [42] to improve generation quality. Examples include Make-A-Video [49], Imagen-Video [23], LVDM [21], Align Your Latent [9], and MagicVideo [69]. Commercial and non-commercial entities have also shown interest in T2V generation, with online services like Gen1 [18], Gen2 [18], and open-source models such as ZeroScope [6], ModelScope [57]. Discord-based servers like Pika-Lab [5] and Morph Studio [4] have demonstrated competitive results.

However, a fair and detailed benchmark for evaluating these methods is still lacking. Existing metrics like FVD [56], IS [45], and CLIP similarity [42] may perform well on previous in-domain T2I generation methods but do not adequately assess alignment with input text, motion quality, and temporal consistency, which are crucial for T2V.

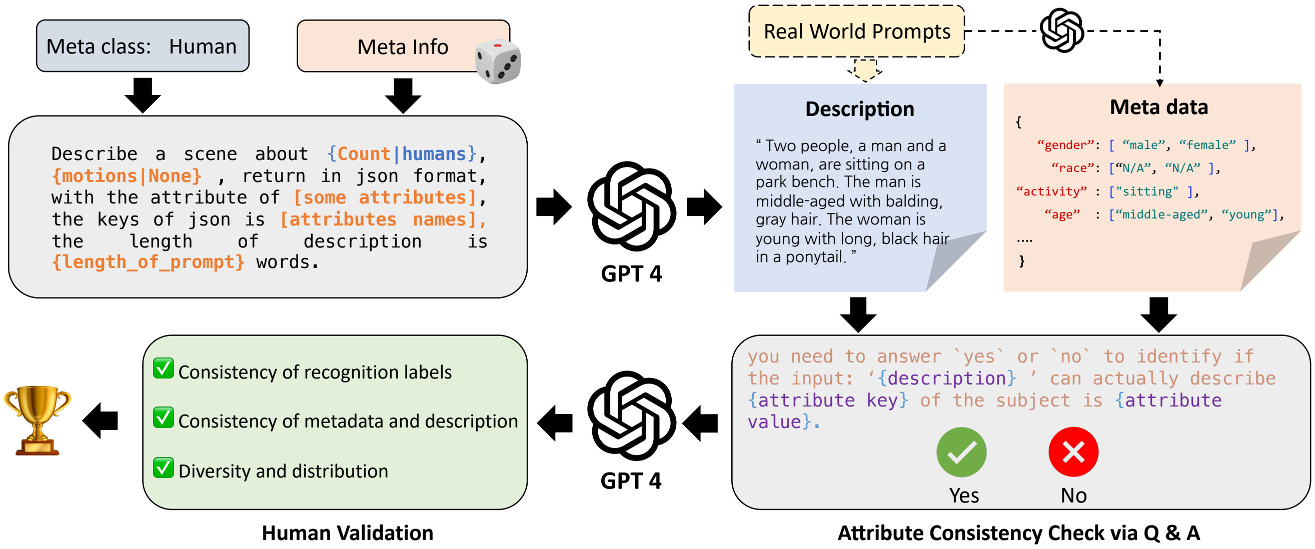
2.2. Evaluations on Large Generative Models

Evaluating the large generative models [37, 40, 43, 54, 55] is a big challenge for both the NLP and vision tasks. For the LLMs, current methods design several metrics in terms of different abilities, question types, and user platform [15, 22, 61, 68, 70]. More details of LLM evaluation and Multimodel LLM evaluation can be found in recent surveys [11, 67]. Similarly, the evaluation of the multi-modal generative model also draws the attention of the researchers [8, 62]. For example, Seed-Bench [33] generates the VQA for multimodal LLM evaluation.

For the models in visual generation tasks, Imagen [44] only evaluates the model via user studies. DALL-Eval [14] assesses the visual reasoning skills and social basis of the T2I model via both user and object detection algorithm [10]. HRS-Bench [7] proposes a holistic and reliable benchmark by generating the prompt with ChatGPT [37] and utilizing 17 metrics to evaluate the 13 skills of the T2I model. TIFA [26] proposes a benchmark utilizing the visual question answering (VQA). However, these methods still work for T2I evaluation or language model evaluation. For T2V evaluation, we further 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 T2V models fairly. To this end, we first collect and analyze large-scale real-world users' prompts. After that, we propose an automatic pipeline to generate a prompt list with high diversity. Since video generation is time-consuming, we collect 700 prompts as our initial version for evaluation with careful annotation. In



Describe a scene about {Count|humans}, {motions|None} , return in json format, with the attribute of [some attributes], the keys of json is [attributes names], the length of description is {length_of_prompt} words.

" Two people, a man and a woman, are sitting on a park bench. The man is middle-aged with balding, gray hair. The woman is young with long, black hair in a ponytail. "

```
{
  "gender": [ "male", "female" ],
    "race": ["N/A", "N/A" ],
"activity" : ["sitting" ],
    "age"  : ["middle-aged", "young"],
....
}
```

you need to answer `yes` or `no` to identify if the input: '{description} ' can actually describe {attribute key} of the subject is {attribute value}.

☑ Consistency of recognition labels

☑ Consistency of metadata and description

☑ Diversity and distribution

Figure 2. 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.

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+	-	-
Show-1	23.10	T2V	576×320	8	✓	4s	10 min	-	-
Hotshot-XL	23.10	T2V	672×384	8	✓	1s	10 s	-	-
VideoCrafter1	23.10	I2V & T2V	1024×576	8	✓	2s	3 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.

this section, we introduce the details of the construction of our benchmark.

Real-World Data Collection. To better understand the types of prompts we should generate, we collect prompts from real-world T2V generation discord users, including the FullJourney [2] and PikaLab [5]. In total, we gather over 600k prompts with corresponding videos and filter them to 200k by removing repeated and meaningless prompts.

Through analyzing the collected data including aspects like prompt length and word frequency, we get to know that most of the prompts contain 3 to 40 words. Besides, we identify four meta-subject classes for T2V generation: `human`, `animal`, `object`, and `landscape`. For each type, we consider the motions and styles of each type, the relationship between the current metaclass and other metaclasses, and the `motion` and `camera motion` to construct the benchmark. We give more details in the supplementary materials.

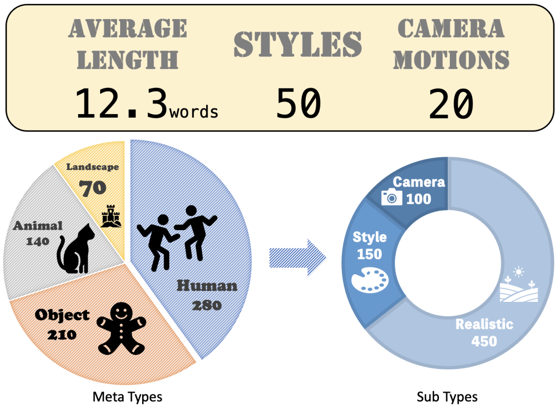


Figure 3. The analysis of the proposed benchmark. Each meta type contains 3 sub-types to increase the generated videos' diversity.

General Recognizable Prompt Generation. Based on the metaclasses identified in the previous step, we generate the recognizable prompts with the help of a LLM and human input. As shown in Fig 2, for each kind of metaclass, we ask GPT-4 [37] to describe the scenes about this metaclass and its attributes with randomly sampled meta information. This way, we get the ground truth for the computer vision models for evaluation. However, we find that GPT-4 is not perfect for this task, as the generated attributes are not very consistent with the generated description. Thus, we involve a self-check in the benchmark building process where we use GPT-4 to identify the similarities between the generated description and each metadata. Finally, we filter the prompts by ourselves to ensure each prompt is correct and meaningful for T2V generation.

In addition to the automatically generated prompts, we also integrate prompts from real-world users and available T2I evaluation prompts, such as DALL-Eval [14] and Draw-Bench [44]. We filter and generate the metadata using GPT-4, choose suitable prompts with corresponding meta-information as shown in Fig. 2, and check the consistency of the meta-information.

Benchmark Overview. Overall, we get over 700 prompts in the metaclasses 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. 3. To increase the diversity of the prompts, our benchmark contains 3 different sub-types, where we have a total of 50 styles and 20 camera motion prompts. We randomly add them in 250 prompts of the whole benchmark. Our benchmark contains an average length of 12.3 words per prompt, which is similar to the real-world prompts we collected.

4. Evaluation Metrics

Different from previous FID [46] based evaluation metrics, we evaluate the T2V models in different aspects, including the visual quality of the generated video, the text-video alignment, the motion quality, and temporal consistency. Below, we give the detailed metrics.

4.1. Overall Video Quality Assessment

We focus on the visual quality of the generated video, which is crucial for user appeal. As distribution-based methods like FVD [56] require ground truth videos, we argue they are unsuitable for general T2V generation cases.

Video Quality Assessment (VQA$_A$, VQA$_T$). We employ the Dover [59] method to assess generated video quality in terms of aesthetics and technicality. The technical rating measures common distortions like noise and artifacts. Dover [59] is trained on a large-scale dataset with labels ranked by real users. We denote the aesthetic and technical scores as VQA$_A$ and VQA$_T$, respectively.

Inception Score (IS). We also use the inception score [45] as a video quality assessment index, following previous T2V generation papers. The inception score evaluates GAN [20] performance using a pre-trained Inception Network [52] on the ImageNet [17] dataset. A higher inception score indicates more diverse generated content.

4.2. Text-Video Alignment

We evaluate the alignment of input text and generated video in various aspects, including global text prompts, content correctness, and specific attributes. The details of each score are as follows.

Text-Video Consistency (CLIP-Score). We use the CLIP-Score to quantify the discrepancy between input text prompts and generated videos. Using the pretrained `ViT-B/32` CLIP model [42] as a feature extractor, we obtain frame-wise image embeddings and text embeddings, and compute their cosine similarity. The overall CLIP-Score is then derived by averaging individual scores across all frames.

Image-Video Consistency (SD-Score). We propose a new metric, SD-Score, to compare the generated quality with frame-wise SD [43], considering that most current video diffusion models are fine-tuned on a base SD with a larger scale dataset. Using SDXL [40], we generate N_1 images $\{d_k\}_{k=1}^{N_1}$ for every prompt and extract visual embeddings in both generated images and video frames. We calculate the embedding similarity between the generated videos and SDXL images, which helps address the concept forgetting problems when fine-tuning the T2I diffusion model to video models. The final SD-Score is calculated as:

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

where x_t^i is the t-th frame of the i-th video, $\mathcal{C}(\cdot,\cdot)$ is the cosine similarity function, $emb(\cdot)$ means CLIP embedding, M is the total number of testing videos, and N is the total number of frames in each video, where $N_1 = 5$.

Text-Text Consistency (BLIP-BLEU). We also consider the evaluation between the generated text descriptions of the video and the input prompt. We utilize BLIP2 [35] for caption generation and use BLEU [39] for evaluation of text alignment:

$$S_{BB} = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{N_2}\sum_{k=1}^{N_2}\mathcal{B}(p^i, l_k^i)), \tag{2}$$

where p^i is the i-th prompt, $\mathcal{B}(\cdot,\cdot)$ is the BLEU similarity scoring function, $\{l_k^i\}_{k=1}^{N_2}$ are BLIP2 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). We employ SAM-Track [13] to analyze the correctness of the video content. 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 presence across videos, calculated as:

$$S_{Det} = \frac{1}{M_1} \sum_{i=1}^{M_1} \left(\frac{1}{K} \sum_{k=1}^{K} \sigma_{t_k}^{i} \right), \tag{3}$$

where M_1 is the number of prompts with objects, K is the number of frames where detection is performed, and $\sigma_{t_k}^{i}$ is the detection result for frame t_k in video i (1 if an object is detected, 0 otherwise). In our approach, we perform detection every $I = 5$ frames. Therefore, $K = \lceil \frac{N}{I} \rceil$.

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

$$S_{Count} = \frac{1}{M_2} \sum_{i=1}^{M_2} \left(1 - \frac{1}{K} \sum_{k=1}^{K} \frac{\left| c_{t_k}^{i} - \hat{c}^{i} \right|}{\hat{c}^{i}} \right), \tag{4}$$

where M_2 is the number of prompts with object counts, $c_{t_k}^{i}$ is the detected object count at frame t_k 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:

$$S_{Color} = \frac{1}{M_3} \sum_{i=1}^{M_3} \left(\frac{1}{K} \sum_{k=1}^{K} s_{t_k}^{i} \right), \tag{5}$$

where M_3 is the number of prompts with object colors and $s_{t_k}^{i}$ is the color accuracy result for frame t_k in video i (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 evaluate the correctness of human faces using DeepFace [47], a popular face analysis toolbox. We calculate the distance between the generated celebrities' faces and real images of the celebrities.

$$S_{CIS} = \frac{1}{M_4} \sum_{i=1}^{M_4} (\frac{1}{N} \sum_{t=1}^{N} (\min_{k \in \{1,\ldots,N_3\}} \mathcal{D}(x_t^i, f_k^i))), \tag{6}$$

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 i-th prompt, and $N_3 = 3$.

Text Recognition (OCR-Score) Another hard case for visual generation is to generate text in the input prompt. To examine the abilities of current T2V models for text generation, we utilize the widely used toolbox PaddleOCR [38] to detect the English text from generated videos. Then, similar to HRS-Bench [7], we calculate Word Error Rate (WER) [30], Normalized Edit Distance (NED) [51], Character Error Rate (CER) [36], and get the average.

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. We use the MMAction2 toolbox [16] and the pre-trained VideoMAE V2 model [58] to infer human actions in generated videos. We take the classification accuracy as our Action-Score, focusing on Kinetics 400 action classes [27].

Average Flow (Flow-Score). We also consider the general motion information of the video. To this end, we use 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 that are hard to be identified 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 5 based on our subjective observation.

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, 41]. 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 [18, 41]. Specifically, we consider the cosine similarity of the embeddings of each two consecutive frames $(emb(x_t), emb(x_{t+1}))$ of the generated videos and then get the averages on each two frames.

Face Consistency. Similar to CLIP-Temp, we evaluate the human identity consistency of the generated videos. Specifically, we select the first frame x_1 as the reference and calculate the cosine similarity of $emb(x_1)$ with $\{emb(x_t)\}_{t=2}^{N}$. Then, we average the similarities as the final score.

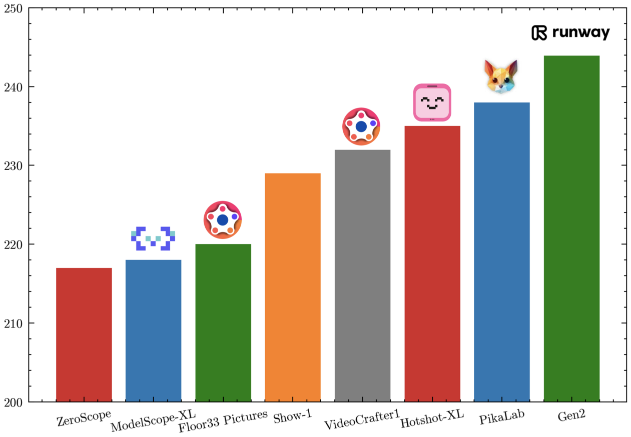


Figure 4. Overall comparison results on our EvalCrafter benchmark.

	Visual Quality	Text-Video Alignment	Motion Quality	Temporal Consistency
ModelScope	53.09 (7)	54.46 (7)	52.47 (7)	57.80 (6)
ZeroScope	53.41 (6)	51.21 (8)	53.61 (4)	58.91 (5)
Floor33 Pictures	58.78 (5)	61.32 (4)	49.16 (8)	50.24 (8)
PikaLab	60.77 (3)	55.80 (6)	55.77 (2)	**65.41 (1)**
Gen2	**62.51 (1)**	60.98 (5)	**56.43 (1)**	64.41 (2)
VideoCrafter1	60.85 (2)	61.95 (2)	53.08 (5)	55.89 (7)
Show-1	52.19 (8)	**62.07 (1)**	53.74 (3)	60.83 (3)
Hotshot-XL	60.38 (4)	61.52 (3)	52.98 (6)	59.96 (4)

Table 2. Human-preference aligned results from four different aspects, with the rank of each aspect in the brackets.

4.5. User Opinion Alignments

Besides the above objective metrics, we evaluate user opinions through studies focusing on five main aspects: (1) *Video Quality*, indicating the quality of the generated video where a higher score shows less blur, noise, or other visual degradation; (2) *Text and Video Alignment*, examining the relationships between the generated video and the input text-prompt, requiring users to evaluate the correctness of generated motions; (3) *Motion Quality*, requiring users to identify the correctness of the generated motions from the video. (4) *Temporal Consistency*, assessing frame-wise consistency, varying from *Motion Quality*, which needs users to give a rank for high-quality movement; (5) *Subjective likeness*, 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 [6], Gen2 [18], Floor33 [1], and PikaLab [5], 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 [40] 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 7 users to give opinions between 1 to 5, where a large value indicates better alignments. The video sequence has been randomly shuffled before being given to users, and we get 8647 feedback scores in total. Finally, after filtering, we keep 1024 most objective and professional scores as illustrated in Fig. 5.

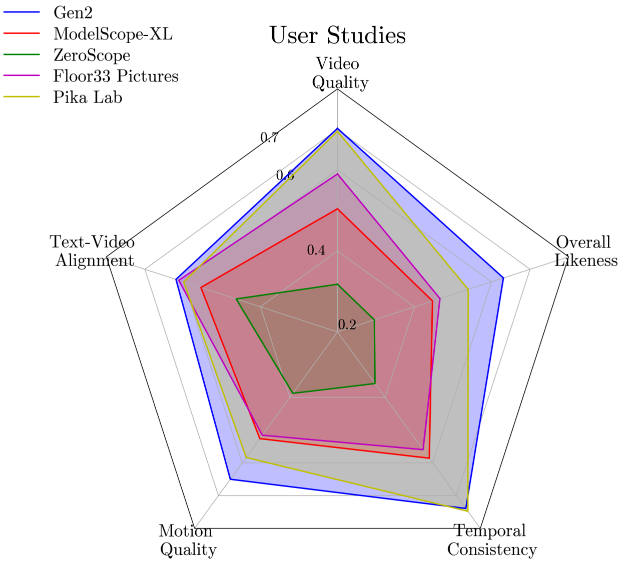


Figure 5. The raw ratings from our user study.

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 specific aspects. Similar to the works of the evaluation of natural language processing [19, 34], we employ a linear regression model to fit the parameters in each dimension. Specifically, we randomly choose 80% samples from four different methods as the fittings samples and left the rest 20% samples to verify the effectiveness of the proposed method. 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 total score to obtain a comprehensive final score.

5. Results

We conduct the evaluation on our benchmark prompts, where each prompt has a metafile for additional information as the answer of evaluation. We then generate the videos using all available high-resolution T2V models, including the ModelScope [57], Floor33 Pictures [1], and ZeroScope [6], Show-1 [65], Hotshot-XL [3], and VideoCrafter1 [12]. We keep all the hyper-parameters, such as classifier-free guid-

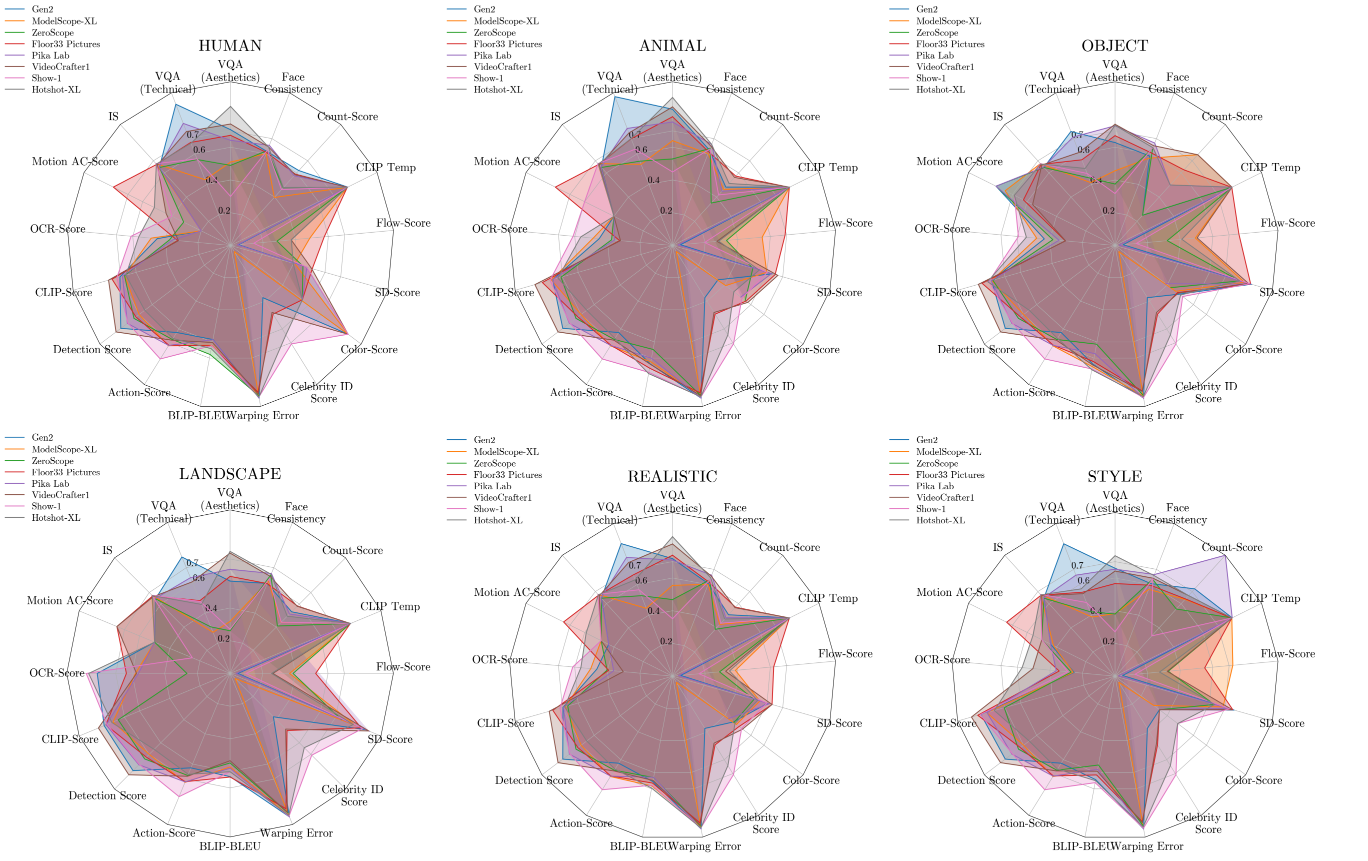


Figure 6. ***Raw results in different aspects.*** We consider 4 main meta types (`animal`, `human`, `landscape`, `object`) to evaluate the performance of the meta types of the generated video, where each type contains several prompts with fine-grained attribute labels. For each prompt, we also consider 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, the Warping Error, Celebrity ID Score, and OCR-Score by 1− so that large values indicate better performance.)

ance, as the default value. For the service-based model, we evaluate the performance of the representative works of Gen2 [18] and PikaLab [5]. They generate at least 512p videos with high-quality watermark-free videos. We run all available models on an NVIDIA A100 for speed comparison. We first show the overall human-aligned results in Fig. 4, with also the different aspects of our benchmark in Table 2, which gives us the final and the main metrics of our benchmark. Finally, as in Fig. 6, we give the results of each method on 4 different meta types (*i.e.*, `animal`, `human`, `landscape`, `object`) and two different types of videos (*i.e.*, `realistic`, `style`) in our benchmark.

5.1. Analysis 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 [28], both of which are non-parametric measures of rank correlation. These coefficients provide insights into the magnitude and direction of the association between our method results and human scores, as listed in Table. 3. From this table, the proposed weighting method shows a better correlation on the unseen 20% samples than directly averaging.

5.2. Findings

Finding #1: Single dimension evaluation is insufficient for nowadays T2V models. Models' rankings in Table. 2 vary significantly across different aspects, emphasizing the importance of a multi-aspect evaluation approach for a comprehensive understanding of model performance.

Finding #2: Meta type evaluation is necessary. As shown in Fig. 6, models perform differently in various meta types, highlighting the importance of evaluating their abilities by meta type. For example, Gen2 [18] behaves better than Floor33 Pictures [1] w.r.t. VQA$_A$ in `human`, `animal`, and `style` videos. Contrarily, it falls behind Floor33 Pictures in `landscape`, `object`, and `realistic` ones.

Finding #3: Users prioritize visual appeal over T2V alignment. As shown in 5, despite Gen2 [18] performing relatively badly in T2V alignment, it surpasses all other models in `Subjective Likeness`. We argue that it is because users prefer videos with better visual appeal like good visual quality and high temporal consistency.

Finding #4: All methods cannot perform camera motion control using prompt. Although some additional hyper-parameters can be set as additional control handles for Gen2 [18] and PikaLab [5], all current models still lack the understanding of open-world prompts like camera motion.

Finding #5: Resolution doesn't correlate much with visual appeal. As shown in Table. 1 and Table. 2, Gen2 [18] and Hotshot-XL [3] have small resolutions but are both competitive in visual quality.

Finding #6: Larger motion amplitude doesn't ensure user preference. In our study, most videos that users are fond of are with slight movements, such as those videos generated by PikaLab [5] and Gen2 [18].

Finding #7: Generating text remains challenging. Most methods struggle to generate high-quality and consistent text from prompts, as evident from OCR-Scores. Raw results of all metrics are given in supplementary materials.

Finding #8: Many models can sometimes generate completely wrong videos. From our study, we find quite a number of failure cases like severe noises and distortion from our baseline models such as ZeroScope [6], ModelScope [57] and Floor33 Pictures [1]. We argue that it could be viewed as a catastrophic forgetting problem [48], as we know many current T2V models are finetuned from base models like SD [43]. We present our detailed qualitative results in supplementary materials.

Finding #9: Effective metrics and not that effective metrics. Metrics like Warp Error, CLIP-Temp, VQA$_T$, and VQA$_A$ seem to perform well as they all have high correlations with human scores shown in Table. 3. However, some metrics are not as good as we think. The Clip-Score especially, which is a widely used metric in previous works [18, 23, 49], only has Spearsman's ρ 6.3 and Kendall's ϕ 4.3 compared to BLIP-BLEU in the same aspects has 26.7 and 19.0. Detailed correlation results can be found in the supplementary materials.

Finding #10: All current models are not satisfactory enough. From our objective evaluation and subjective observation, we argue that T2V models nowadays still have lots to improve. Even for the best model in our evaluation, Gen2 [18] also has limitations like struggling with complex scenes, instruction following, and entity details.

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 700 prompts as the benchmark, where the real-world situation is very complicated. More prompts will show a more general 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 fewer 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.

Aspects	Methods	Spearsman's ρ	Kendall's ϕ
Visual Quality	VQA$_A$	42.1	30.5
	VQA$_T$	53.6	39.1
	Avg.	55.0	41.0
	Ours	**55.4**	**41.1**
Motion Amplitude	Motion AC	-22.1	-16.4
	Flow-Score	-43.3	-30.1
	Avg.	-38.2	-27.7
	Ours	**45.0**	**32.4**
Temporal Consistency	CLIP-Temp	49.8	35.7
	Warping Error	69.0	51.7
	Avg.	54.4	38.9
	Ours	**56.7**	**41.5**
TV Alignment	CLIP-Score	6.3	4.3
	BLIP-BLEU	26.7	19.0
	Avg.	31.9	**22.7**
	Ours	**32.3**	22.5

Table 3. ***Correlation Analysis.*** Correlations between some objective metrics and human judgment on text-to-video generations. We use Spearsman's ρ and Kendall's ϕ for correlation calculation.

6. Conclusion

Exploring the capabilities of large generative models is crucial for improving model design and utilization. In this paper, we take the first step towards evaluating large, high-quality T2V models by constructing a comprehensive prompt benchmark for T2V assessment. We also provide several objective evaluation metrics to measure T2V model performance concerning video quality, text-video alignment, temporal consistency, and motion quality. Furthermore, we conduct human alignment to correlate user scores with objective metrics, resulting in accurate evaluation metrics for T2V methods. Our experiments demonstrate that the proposed methods effectively align with user opinions, thus providing a reliable assessment of T2V approaches. We believe this comprehensive evaluation benchmark will serve as a foundation and foster development for future research.

Acknowledgments

Xuebo Liu was sponsored by CCF-Tencent Rhino-Bird Open Research Fund. We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

[1] Floor33 pictures discord server. https://www.morphstudio.com/. Accessed: 2023-08-30. 6, 7, 8

[2] Fulljourney discord server. https://www.fulljourney.ai/. Accessed: 2023-08-30. 3

[3] Hotshot-xl. https://huggingface.co/hotshotco/Hotshot-XL. Accessed: 2023-10-11. 6, 8

[4] Morph studio discord server. https://www.morphstudio.com/. Accessed: 2023-08-30. 2

[5] Pika Lab discord server. https://www.pika.art/. Accessed: 2023-08-30. 1, 2, 3, 6, 7, 8

[6] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed: 2023-08-30. 2, 6, 8

[7] 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. 2, 5

[8] 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. 2

[9] 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

[10] 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. 2

[11] 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. 2

[12] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 6

[13] 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. 4

[14] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. 2, 4

[15] 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

[16] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 5

[17] 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. 4

[18] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1, 2, 5, 6, 7, 8

[19] 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. 6

[20] 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, 4

[21] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. 1, 2

[22] 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

[23] 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, 8

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[25] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2

[26] 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

[27] 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. 5

[28] Maurice George Kendall. Rank correlation methods. 1948. 7

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[30] Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002. 5

[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. 5

[32] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020. 5

[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

[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. 6

[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. 4

[36] 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. 5

[37] OpenAI. Gpt-4 technical report, 2023. 1, 2, 4

[38] PaddlePaddle. Paddleocr. https://github.com/PaddlePaddle/PaddleOCR, 2013. 5

[39] 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. 4

[40] 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, 4, 6

[41] 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. 5

[42] 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, 4

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 4, 8

[44] 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. 2, 4

[45] 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, 4

[46] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. 4

[47] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 5

[48] Chenze Shao and Yang Feng. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. *arXiv preprint arXiv:2203.03910*, 2022. 8

[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, 8

[50] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. 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. 5

[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. 4

[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. 5

[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, Prajjwal 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, 6, 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. 5

[59] 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

[60] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 2

[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. 2

[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. 7

[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. *arXiv preprint arXiv:2309.15818*, 2023. 6

[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. 2

[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