VBench: Comprehensive Benchmark Suite for Video Generative Models

Ziqi Huang[1*] Yinan He[2*] Jiashuo Yu[2*] Fan Zhang[2*] Chenyang Si[1] Yuming Jiang[1]
Yuanhan Zhang[1] Tianxing Wu[1] Qingyang Jin[1] Nattapol Chanpaisit[1]
Yaohui Wang[2] Xinyuan Chen[2] Limin Wang[4,2] Dahua Lin[2,3✉] Yu Qiao[2✉] Ziwei Liu[1✉]

[1]S-Lab, Nanyang Technological University [2]Shanghai Artificial Intelligence Laboratory
[3]The Chinese University of Hong Kong [4]Nanjing University

https://vchitect.github.io/VBench-project/

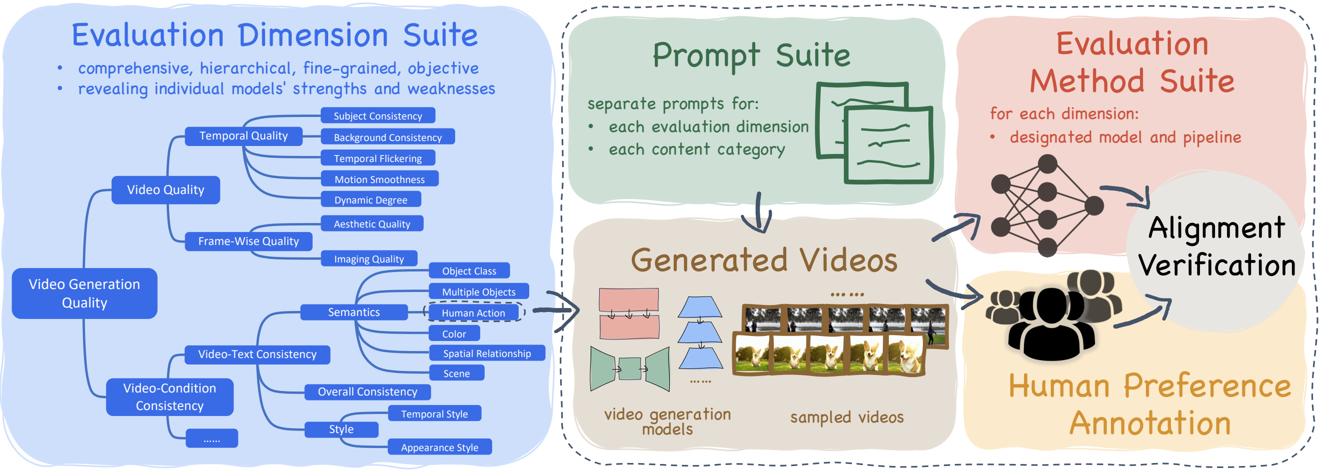


Figure 1. **Overview of VBench.** We propose VBench, a comprehensive benchmark suite for video generative models. We design a comprehensive and hierarchical **Evaluation Dimension Suite** to decompose "video generation quality" into multiple well-defined dimensions to facilitate fine-grained and objective evaluation. For each dimension and each content category, we carefully design a **Prompt Suite** as test cases, and sample **Generated Videos** from a set of video generation models. For each evaluation dimension, we specifically design an **Evaluation Method Suite**, which uses a carefully crafted method or designated pipeline for automatic objective evaluation. We also conduct **Human Preference Annotation** for the generated videos for each dimension and show that VBench evaluation results are **well aligned with human perceptions**. VBench can provide valuable insights from multiple perspectives.

Abstract

*Video generation has witnessed significant advancements, yet evaluating these models remains a challenge. A comprehensive evaluation benchmark for video generation is indispensable for two reasons: 1) Existing metrics do not fully align with human perceptions; 2) An ideal evaluation system should provide insights to inform future developments of video generation. To this end, we present **VBench**, a comprehensive benchmark suite that dissects "video generation quality" into specific, hierarchical, and disentangled dimensions, each with tailored prompts and evaluation methods. VBench has three appealing properties: **1) Comprehensive Dimensions:** VBench comprises 16 dimensions in video generation (*e.g.*, subject identity inconsistency, motion smoothness, temporal flickering, and spatial relationship,* etc.*). The evaluation metrics with fine-grained levels reveal individual models' strengths and weaknesses. **2) Human Alignment:** We also provide a dataset of human preference annotations to validate our benchmarks' alignment with human perception, for each evaluation dimension respectively. **3) Valuable Insights:** We look into current models' ability across various evaluation dimensions, and various content types. We also investigate the gaps between video and image generation models. We will open-source VBench, including all prompts, evaluation methods, generated videos, and human preference annotations, and also include more video generation models in VBench to drive forward the field of video generation.*

[*] equal contributions. ✉corresponding authors. Code is available

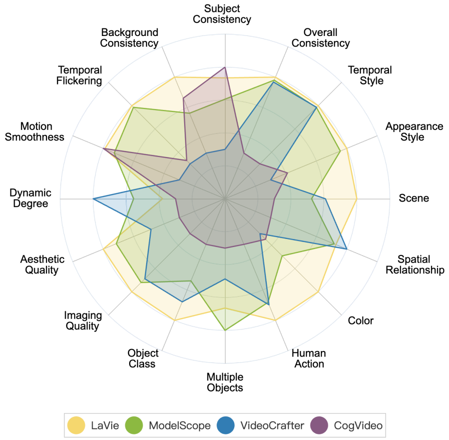


Figure 2. **VBench Evaluation Results of Video Generative Models.** We visualize the evaluation results of four video generation models in 16 VBench dimensions. We normalize the results per dimension for clearer comparisons. For comprehensive numerical results, please refer to Table 1.

1. Introduction

Image generation models have made rapid progress in the past few years, such as Variational Autoencoders (VAEs) [51], Generative Adversarial Networks (GANs) [6, 19, 20, 24, 39, 41, 44–47, 67], vector quantized (VQ) based approaches [16, 40, 85], and diffusion models [31, 78, 79]. This fuels recent explorations in video generation [3, 5, 10, 21, 23, 27, 29, 34, 52, 62, 66, 77, 86, 87, 89, 91, 101, 107–109], which goes beyond static imagery and models the dynamics and kinematics of real-world scenes. With the growth of video generation models, there arises a critical need for effective evaluation methods. The evaluation should be able to accurately reflect human perception of generated videos, providing reliable measures of a model's performance. Additionally, it should reflect each model's specific strengths and weaknesses, offering insights that inform the data, training, and architectural choices of future video generation models.

However, existing metrics for video generation such as Inception Score (IS) [76], Fréchet inception distance (FID) [30], Fréchet Video Distance (FVD) [83, 84], and CLIPSIM [73] are inconsistent with human judgement [15, 69]. Meanwhile, the Video Quality Assessment (VQA) methods [55, 82, 93–99] are primarily designed for real videos, thereby neglecting the unique challenges posed by generative models, such as artifacts in synthesized videos. Hence, there is a pressing need for an evaluation framework that aligns closely with human perception, and specifically designed for the characteristics of video generation models.

To this end, we introduce **VBench**, a comprehensive benchmark suite for evaluating video generation model performance. VBench has three appealing properties: 1) comprehensive evaluation dimensions, 2) human alignment, and 3) valuable insights.

First, our framework includes an *evaluation dimension suite* that employs a hierarchical and disentangled approach to the decomposition of "video generation quality". This suite systematically breaks down the evaluation into two primary dimensions at a coarse level: *Video Quality* and *Video-Condition Consistency*. Each of these dimensions is further subdivided into more granular criteria. This hierarchical separation ensures that each dimension isolates and evaluates a single aspect of video quality, without interference from other variables, as illustrated in Figure 1. Recognizing video generation's unique challenges, we have tailored evaluation dimensions to its specific characteristics. For example, in terms of *Video Quality*, maintaining consistent subject identity (*e.g.*, a teddy bear) in generated videos is crucial, and is a problem rarely encountered in real-world videos. Additionally, *Video-Condition Consistency* is vital for conditional video generation tasks, requiring its dedicated evaluation criteria. For each evaluation dimension, we carefully prepared around 100 text prompts as test cases for text-to-video (T2V) generation, and devised specialized evaluation methods tailored to each dimension. In addition to multi-dimensional evaluations, we also assess T2V models across *diverse content categories*. We organized prompt suites for eight distinct types, such as animal, architecture, human, and scenery, allowing for a separate evaluation within each category. This exploration reveals variable competencies in T2V generation across different content types, highlighting areas of proficiency and those requiring further enhancement.

Second, we systematically demonstrate that our evaluation method suite *is closely aligned with human perception* in every fine-grained evaluation dimension. We collected human preference annotations for each dimension. Specifically, we use various T2V models to sample videos from our prompt suites. Then given two videos sampled from the same prompt, we ask human annotators to indicate preferences according to each VBench dimension respectively. We show that VBench evaluations highly correlate with human preferences. Additionally, the *human preference annotations* can be utilized for multiple purposes, such as fine-tuning generation or evaluation models to enhance alignment with human perceptions. For instance, we utilize the annotations to implement Instruction Tuning within a Visual-Language Model (VLM), enhancing its T2V evaluation alignment with human preferences.

Third, VBench's multi-dimensional and multi-categorical approach can provide *valuable insights* to the video generation community. Our multi-dimensional system enables detailed feedback on the strengths and weaknesses of video generation models across various ability aspects. This approach not only ensures a comprehensive evaluation of existing models but also provides valuable insights into the training of advanced video generation models, guiding architectural and data choices for improved video generation outcomes. Additionally, VBench can be readily applied to evaluate image generation models, and thus we investigate the disparities between video and image generation models. In Section 5, we discuss in detail on various observations and insights drawn from VBench evaluations.

We are open-sourcing **VBench**, including its *evaluation dimension suite*, *evaluation method suite*, *prompt suite*, *generated videos*, and the dataset of *human preference annotations*. We also encourage more video generation models to participate in the **VBench** challenge.

2. Related Works

Video Generative Models. Recently, diffusion models [14, 31, 78, 79] have achieved significant progress in image synthesis [25, 37, 38, 68, 71, 74, 75], and enabled a line of works towards video generation [5, 9, 22, 26, 28, 29, 32, 33, 42, 50, 62, 63, 77, 87, 91, 102, 103, 109, 113]. Many recent diffusion-based works [29, 62, 87, 91] are text-to-video (T2V) models. Other guidance modalities are also available, including image-to-video [11, 13, 17, 106], video-to-video [8, 59, 70, 72, 105], and a variety of control maps [12, 43, 50, 64, 90, 110, 111] such as pose, depth, and sketch. The boom of video generation models requires a comprehensive evaluation system to inform their current capabilities and guide future developments, and VBench takes the initiative in providing a comprehensive benchmark suite for fine-grained and human-aligned evaluation.

Evaluation of Visual Generative Models. Existing video generation models typically use metrics like Inception Score (IS) [76], Fréchet inception distance (FID) [30], Fréchet Video Distance (FVD) [83], and CLIPSIM [73] for evaluation. The UCF-101 [80] dataset's class labels often serve as text prompts for IS, FID, and FVD, whereas MSR-VTT [104]'s human-labeled video captions are used for CLIPSIM. Despite covering various real-world scenarios, these prompts lack diversity and specificity, limiting accurate and fine-grained evaluation of video generation. For text-to-image (T2I) models, several benchmarks [2, 4, 35, 54, 65, 75, 88] are proposed to assess various capabilities like compositionality [35] and editing ability [4, 88]. However, video generative models still lack comprehensive evaluation benchmarks for detailed and human-aligned feedback. Our work differs from concurrent research [60, 61] in three key ways: 1) We have created 16 distinct evaluation dimensions, each with specialized prompts for precise assessment; 2) We have empirically validated that every dimension aligns closely with human perception; 3) Our multi-dimensional and multi-categorical evaluation offers valuable and comprehensive insights into video generation.

3. VBench Suite

In this section, we introduce the main components of VBench. In Section 3.1, we present our rationale for designing the 16 evaluation dimensions, as well as each dimension's definition and evaluation method. We then elaborate on the prompt suites we use in Section 3.2. To validate VBench's alignment with human perception, we conduct human preference annotation for each dimension (see Section 3.3). The experiments and the insights drawn from VBench will be detailed in Section 4 and Section 5.

3.1. Evaluation Dimension Suite

We first introduce our evaluation dimensions and their corresponding evaluation methods.

Existing evaluation metrics like FVD [83] often conclude video generation model performance to a single number. This oversimplifies the evaluation and has several risks. First, a single number can obscure an individual model's strengths and weaknesses, and it fails to provide insights into specific areas where a model excels or underperforms. This makes it challenging to derive insights for future architectural and training designs based on single-valued metrics. Second, the notion of "high-quality video generation" is complex and multifaceted, with individuals prioritizing different video attributes based on the intended application. For instance, some may prioritize the absence of temporal flickering, while others may consider fidelity to the text prompt as the most significant, with less emphasis on flickering. Therefore, in contrast with performing single-valued evaluations of video generation quality, we propose a disaggregated approach by decomposing the brand notion of "video generation performance" into multiple discrete dimensions for fine-grained evaluation.

Specifically, we break "video generation quality" down into 16 disentangled dimensions in a top-down manner, with each evaluation dimension assessing one aspect of video generation quality. On the top level, we evaluate T2V performance from two broad perspectives: **1) Video Quality** — *"Without considering alignment with the text prompt, does the video alone look good?"*, which focuses on the perceptual quality of the synthesized video, and does not consider the input condition (*e.g.*, text prompt), and **2) Video-Condition Consistency** — *"Is the video consistent with what the user wants to generate?"*, which focuses on whether the synthesized video is consistent with the guiding condition that the user provides (*e.g.*, the text prompt for T2V generation). Under both *"Video Quality"* and *"Video-Condition Consistency"*, we further break the coarse-grained dimensions into more fine-grained dimensions, as shown in Figure 1. We briefly introduce each dimension here. *Please refer to the Supplementary File for the detailed definition and evaluation method of each dimension.*

3.1.1 Video Quality

We split *"Video Quality"* into two disentangled aspects, *"Temporal Quality"* and *"Frame-Wise Quality"*, where the former only considers the cross-frame consistency and dynamics, and the latter only considers the quality of each individual frame without taking temporal quality into concern. For *"Temporal Quality"*, we further devise five evaluation dimensions, where each focusing on a different aspect of temporal quality.

Temporal Quality - Subject Consistency. For a subject (*e.g.*, a person, a car, or a cat) in the video, we assess whether its appearance remains consistent throughout the whole video. To this end, we calculate the DINO [7] feature similarity across frames.

Temporal Quality - Background Consistency. We evaluate the temporal consistency of the background scenes by calculating CLIP [73] feature similarity across frames.

Temporal Quality - Temporal Flickering. Generated videos can exhibit imperfect temporal consistency at *local and high-frequency details*. We take static frames and compute the mean absolute difference across frames.

Temporal Quality - Motion Smoothness. *Subject Consistency* and *Background Consistency* focus on temporal consistency of the "look" instead of the smoothness of "movement and motion". We believe it is important to evaluate whether the motion in the generated video is smooth, and follows the physical law of the real world. We utilize the motion priors in the video frame interpolation model [58] to evaluate the smoothness of generated motions.

Temporal Quality - Dynamic Degree. Since a completely static video can score well in the aforementioned temporal quality dimensions, it is important to also evaluate the degree of dynamics (*i.e.*, whether it contains large motions) generated by each model. We use RAFT [81] to estimate the degree of dynamics in synthesized videos.

Frame-Wise Quality - Aesthetic Quality. We evaluate the artistic and beauty value perceived by humans towards each video frame using the LAION aesthetic predictor [53]. It can reflect aesthetic aspects such as the layout, the richness and harmony of colors, the photo-realism, naturalness, and artistic quality of the video frames.

Frame-Wise Quality - Imaging Quality. Imaging quality refers to the distortion *(e.g., over-exposure, noise, blur)* presented in the generated frames, and we evaluate it using the MUSIQ [49] image quality predictor trained on the SPAQ [18] dataset.

3.1.2 Video-Condition Consistency

We mainly dissect *"Video-Condition Consistency"* into *"Semantics"* (*i.e.*, the type of the entities and their attributes) and *"Style"* (*i.e.*, whether the generated video is consistent with user-requested style), with each decomposed into more fine-grained dimensions.

Semantics - Object Class. We use GRiT [100] to detect the success rate of generating the specific class of objects depicted in the text prompt.

Semantics - Multiple Objects. Other than generating a single object of a particular class, the ability to compose multiple objects from different classes in the same frame is also an essential ability in video generation. We detect the success rate of generating all the objects specified in the text prompt within each video frame.

Semantics - Human Action. Human action is an important aspect in human-centric video generation. We apply UMT [57] to evaluate whether human subjects in generated videos can accurately execute the specific actions mentioned in the text prompts.

Semantics - Color. To evaluate whether synthesized object colors align with the text prompt, we use GRiT [100] for color captioning and comparison with expected colors.

Semantics - Spatial Relationship. Other than classes and attributes of synthesized objects, we also evaluate whether their spatial relationship follows what is specified by the text prompt. We focus on four primary types of spatial relationships, and perform rule-based evaluation similar to [35].

Semantics - Scene. We need to evaluate whether the synthesized video is consistent with the intended scene described by the text prompt. For example, when prompted "ocean", the generated video should be "ocean" instead of "river". We use Tag2Text [36] to caption the generated scenes, and then check its correspondence with scene descriptions in the text prompt.

Style - Appearance Style. Apart from semantics consistency with the text prompt, another important pillar in video-condition consistency is *style*. There are many styles that alter the look, color, and texture of synthesized video frames, such as "oil painting style", "black and white style", "watercolor painting style", "cyberpunk style", "black and white" *etc*. We calculate the CLIP [73] feature similarity between synthesized frames and these style descriptions.

Style - Temporal Style. Apart from appearance styles, videos also have temporal styles like various camera motions. We use ViCLIP [92] to calculate the video feature and the temporal style description feature similarity to reflect temporal style consistency.

Overall Consistency. We further use overall video-text consistency computed by ViCLIP [92] on general text prompts as an aiding metric to reflect both semantics and style consistency.

3.2. Prompt Suite

The sampling procedure of current diffusion-based video generation models [29, 87, 91] is time-consuming (*e.g.*, 90 seconds per video for LaVie [91], and more than 2 minutes

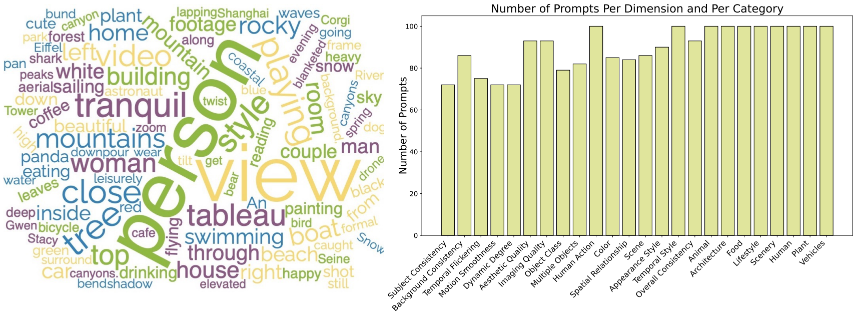
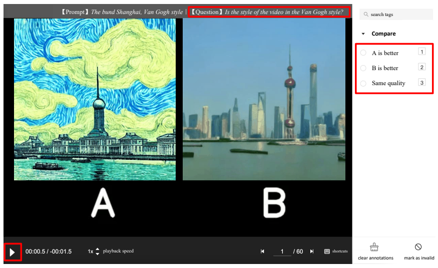


Figure 3. **Prompt Suite Statistics.** The two graphs provide an overview of our prompt suites. ***Left:*** the word cloud to visualize word distribution of our prompt suites. ***Right:*** the number of prompts across different evaluation dimensions and different content categories.



【Prompt】*The bund Shanghai, Van Gogh style* | 【Question】*Is the style of the video in the Van Gogh style?*

○ A is better 1
○ B is better 2
○ Same quality 3

Figure 4. **Interface for Human Preference Annotation**. ***Top:*** prompt and question. ***Right:*** choices that annotators can make. ***Bottom left:*** control for stop and playback.

per video for CogVideo [34]). Therefore, we need to control the amount of test cases for efficient evaluation. Meanwhile, we need to maintain the diversity and comprehensiveness of our prompt suite, so we design compact yet representative prompts in terms of both the evaluation dimensions and the content categories. We visualize our prompt suite distributions in Figure 3.

Prompt Suite per Dimension. For each VBench evaluation dimension, we carefully designed a suite of around 100 prompts as test cases. The prompt suite is carefully curated to probe the specific ability relevant to the dimension tested. For example, for the *"Subject Consistency"* dimension which aims to evaluate the consistency of subjects' appearances throughout the video, we ensure every prompt has a movable subject (*e.g.*, animals or vehicles) performing non-static actions, where their consistency might be compromised due to inconsistency introduced by their movements or changing locations. In *"Object Class"* dimension, we ensure the existence of a specific class of object in every prompt. For *"Human Action"*, each test prompt contains a human subject performing a well-defined action from the Kinetics-400 dataset [48], where 100 representative actions are selected with minimal semantic overlaps among themselves. Please refer to the Supplementary File for the design rationale of the prompt suite for all 16 dimensions.

Prompt Suite per Category. When designing prompts for each dimension, the focus was to showcase models' ability in that specific dimension. We further incorporate prompt suites for eight content categories to provide insights into the performance across varied content types. To this end, we prepare a collection of human-curated prompts from the Internet and divide them into 8 distinctive categories following YouTube's categorization. Subsequently, we feed both the category labels and prompts into a Large Language Model (LLM) [112] (see more implementation details in Supplementary File), obtaining multi-label outputs for each caption. We select 800 prompts and manually clean their labels to serve as per-category prompt suites. Finally, we obtain 100 prompts for each of these eight categories: `Animal`, `Architecture`, `Food`, `Human`, `Lifestyle`, `Plant`, `Scenery`, and `Vehicles`.

3.3. Human Preference Annotation

We perform human preference labeling on massive generated videos. The primary goal is to validate *VBench evaluation's alignment with human perception in each of the 16 evaluation dimensions*, and the verification results will be detailed in Section 4.2. We also show that our human preference annotations can be useful in future tasks of finetuning generation and evaluation models to enhance alignment with human perceptions.

Data Preparation. Given a text prompt p_i, and four video generation models to be evaluated $\{A, B, C, D\}$, we use each model to generate a video, forming a "group" of videos $G_{i,j} = \{V_{i,A,j}, V_{i,B,j}, V_{i,C,j}, V_{i,D,j}\}$. For each prompt p_i, we sample five such groups of videos $\{G_{i,0}, G_{i,1}, G_{i,2}, G_{i,3}, G_{i,4}\}$. For each group, we pair the videos up in pair-wise combinations, yielding six pairs: (V_A, V_B), (V_A, V_C), (V_A, V_D), (V_B, V_C), (V_B, V_D), (V_C, V_D), and ask human annotators to indicate their preferred video for each pair. Within the VBench evaluation framework, a prompt suite of N prompts produces $N \times 5 \times 6$ pairwise video comparisons. The video order within each pair is randomized to ensure unbiased annotation.

Human Labeling Rules. Specifically, the human annotators are asked to only consider the specific evaluation dimension of interest and select the preferred video. For example, in Figure 4, for the *Appearance Style* dimension, the question is *"Is the style of the video in the Van Gogh style?"*, and human annotators are instructed to only focus on whether the generated video's style belongs to the Van Gogh style and should not consider other quality aspects of the generated video, such as potential issues like the degree of temporal flickering. In the example in this figure, video A resembles the Van Gogh better than video B, and the annotator is expected to select "A is better". For every dimension, we carefully prepare instructions and train the human annotators to understand the definition of the dimension, and perform multiple quality assurance protocols via a pre-labeling trial, and two rounds of post-labeling checks

Annotations for VLM Tuning. We map VBench evaluation scores from various dimensions to the scale of 0-10 and combine them with human preference annotations to

Table 1. **VBench Evaluation Results per Dimension.** This table compares the performance of four video generation models across each of the 16 VBench dimensions. A higher score indicates relatively better performance for a particular dimension. We also provide two specially built baselines, *i.e*., Empirical Min and Max (the approximated achievable min and max scores for each dimension), as references.

Models	**Subject Consistency**	**Background Consistency**	**Temporal Flickering**	**Motion Smoothness**	**Dynamic Degree**	**Aesthetic Quality**	**Imaging Quality**	**Object Class**
LaVie [91]	91.41%	**97.47%**	**98.30%**	96.38%	49.72%	**54.94%**	**61.90%**	**91.82%**
ModelScope [62, 87]	89.87%	95.29%	98.28%	95.79%	66.39%	52.06%	58.57%	82.25%
VideoCrafter [29]	86.24%	92.88%	97.60%	91.79%	**89.72%**	44.41%	57.22%	87.34%
CogVideo [34]	**92.19%**	96.20%	97.64%	**96.47%**	42.22%	38.18%	41.03%	73.40%
Empirical Min	14.62%	26.15%	62.93%	70.60%	0.00%	0.00%	0.00%	0.00%
Empirical Max	100.00%	100.00%	100.00%	99.75%	100.00%	100.00%	100.00%	100.00%

Models	**Multiple Objects**	**Human Action**	**Color**	**Spatial Relationship**	**Scene**	**Appearance Style**	**Temporal Style**	**Overall Consistency**
LaVie [91]	33.32%	**96.80%**	**86.39%**	34.09%	**52.69%**	**23.56%**	**25.93%**	**26.41%**
ModelScope [62, 87]	**38.98%**	92.40%	81.72%	33.68%	39.26%	23.39%	25.37%	25.67%
VideoCrafter [29]	25.93%	93.00%	78.84%	**36.74%**	43.36%	21.57%	25.42%	25.21%
CogVideo [34]	18.11%	78.20%	79.57%	18.24%	28.24%	22.01%	7.80%	7.70%
Empirical Min	0.00%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	0.00%
Empirical Max	100.00%	100.00%	100.00%	100.00%	82.22%	28.55%	36.40%	36.40%

form the instruction data, which is then used to fine-tune the pre-trained VideoChat [56] model to demonstrate improved evaluation capabilities. For implementation details and tuning results, please refer to the Supplementary File.

4. Experiments

We adopt the video generation models LaVie [91], ModelScope [62, 87], VideoCrafter [29], and CogVideo [34] for VBench evaluation, and more will be added as they become open-sourced. Details of the models and sampling procedures are in the Supplementary File.

4.1. Per-Dimension Evaluation

For every dimension, we calculate the VBench scores using the evaluation method suite described in Section 3.1, and show the results using Figure 2 and Table 1. We additionally designed three reference baselines, namely *Empirical Max*, *Empirical Min*, and *WebVid-Avg*. The first two approximate the maximum / minimum scores that videos might be able to achieve, and *WebVid-Avg* reflects the WebVid-10M [1] dataset quality in each VBench dimension.

Empirical Max. For most dimensions, to approximate the maximum achievable values, we first retrieve WebVid-10M [1] videos according to our prompt suites. We use CLIP [73] to extract text features of both WebVid-10M's captions and our prompts. For each prompt, we retrieve the top-5 WebVid-10M videos according to text feature similarity with the given prompt. Given that the generated videos are usually 2 seconds in length, we randomly select a 2-second segment from each retrieved video and sample frames at 8 frames per second (FPS). For each dimension, we use the retrieved videos according to its prompt suite and report the highest-scoring video's result as *Empirical Max*.

Empirical Min. To approximate the minimum achievable values, we use randomly generated 2-second Gaussian noise clips to calculate results for the *"Video-Condition Consistency"* dimensions. For most *"Video Quality"* dimensions, we select frames from real videos and design frame concatenation for each dimension, approximating the minimum score achievable for each VBench dimension.

WebVid-Avg. Similar to *Empirical Max*, we compute the average for each dimension on retrieved WebVid-10M [1] videos. This baseline could reflect the average per-dimension quality of the commonly used video generation training dataset WebVid-10M, and provide a reference for model performances. The comparison against *WebVid-Avg* and *Empirical Max* is visualized in Figure 6 (b).

4.2. Validating Human Alignment of VBench

To validate that our evaluation method can faithfully reflect human perception, we performed a large-scale human annotation for each dimension, as mentioned in Section 3.3. We show the correlation between VBench evaluation results and human preference annotations in Figure 5.

Win Ratio. Given the human labels, we calculate the win ratio of each model. During pairwise comparisons, if a model's video is selected as better, then the model scores 1 and the other model scores 0. If there is a tie, then both models score 0.5. For each model, the win ratio is calculated as the total score divided by the total number of pairwise comparisons participated.

Per-Dimension Evaluation. For each evaluation dimension, we calculate the model win ratio based on (1) VBench evaluation results, and (2) human annotation results, respectively, and compute their correlations, as shown in Figure 5. We observe that *VBench's per-dimension evaluation results are highly correlated with human preference annotations*.

4.3. Per-Category Evaluation

We evaluate the T2V models across eight different content categories, by generating videos based on *Prompt Suite per Category* described in Section 3.2, and then calculating their performance across different evaluation dimensions. Figure 7 visualizes the evaluation results of each model in terms of the eight content categories.

4.4. Video Generation V.S. Image Generation

We conduct a comparative analysis of the frame-wise generation capability exhibited by text-to-video (T2V) models and text-to-image (T2I) models with two primary objectives: first, to assess the extent to which T2V models

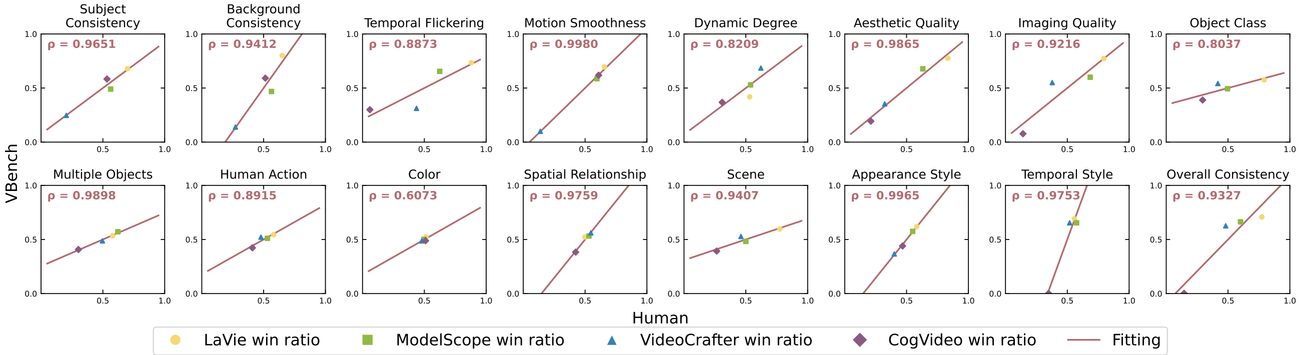
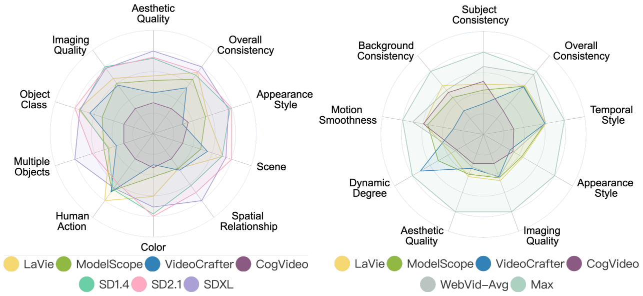


Figure 5. **Validate VBench's Human Alignment.** Our experiments show that ***VBench evaluations across all dimensions closely match human perceptions.*** Each plot shows the alignment verification result of a specific VBench dimension. In each plot, a dot represents the human preference win ratio (horizontal axis) and VBench evaluation win ratio (vertical axis) for a particular video generation model. We linearly fit a straight line to visualize the correlation, and calculate the Spearman's correlation coefficient (ρ) for each dimension.



(a) T2V vs. T2I

(b) T2V vs. WebVid-Avg & Max

Figure 6. **More Comparisons of Video Generation Models with Other Models and Baselines.** We use VBench to evaluate other models and baselines for further comparative analysis of T2V models. **(a)** Comparison with text-to-image (T2I) generation models. **(b)** Comparison with *WebVid-Avg* and *Empirical Max* baselines. See the Supplementary File for comprehensive numerical results and details on normalization methods.

have successfully inherited the frame-wise generative capability of the T2I models; and second, to investigate the frame-wise generation capability gap between existing T2I and T2V models. As an initial exploration into this problem, we compare video generation models with three image generation models, namely Stable Diffusion (SD) 1.4 [74], SD2.1 [74], and SDXL [71]. We choose 10 VBench dimensions that can encompass frame-wise generation capabilities, and sample frames from all the image and video generation models according to *Prompt Suite per Evaluation Dimension* described in Section 3.2. Figure 6 (a) visualizes the evaluation results of T2V versus T2I models.

5. Insights and Discussions

In this section, we discuss the observations and insights we draw from our comprehensive evaluation experiments.

· **Trade-off across Ability Dimensions.** We have noticed a trade-off in video generation models between ***1)*** temporal consistency (*Subject Consistency*, *Background Consistency*, *Temporal Flickering*, *Motion Smoothness*) and ***2)*** *Dynamic Degree*. Models strong in temporal consistency often have a lower *Dynamic Degree*, as these two aspects are somewhat complementary (see Figure 2 and Table 1). For example, LaVie excels in *Background Consistency* and *Temporal Flickering* but has a low *Dynamic Degree*, probably because generating relatively static scenes can "cheat" to get high temporal consistency scores. Conversely, VideoCrafter shows a high *Dynamic Degree* but suffers from poor performance in all temporal consistency dimensions. This trend highlights the current challenge for models to achieve temporal consistency with dynamic content of large motions. Future research should focus on enhancing both aspects simultaneously, as improving only one might indicate compromising the other.

· **Uncovering Hidden Potential of T2V Models in Specific Content Categories.** Our analysis reveals that the capabilities of some models vary significantly across different content types. For instance, for *Aesthetic Quality*, CogVideo scores well for `Food` (see Figure 7 rightmost chart), whereas it underperforms in others like `Animal` and `Vehicles`. The average results across various prompts might suggest a lower overall *"Aesthetic Quality"* (as seen in Figure 2), but CogVideo demonstrates relatively strong aesthetics in at least the `Food` category. This suggests that with tailored training data and strategies, CogVideo could potentially match other models in aesthetics by improving such ability in other content types. Therefore, we recommend *evaluating video generation models not just based on ability dimensions but also considering specific content categories to uncover their hidden potential.*

· **Bottleneck in Temporally Complex Categories Affecting Spatial and Temporal Performance.** For spatially complex categories (*e.g.*, `Animal`, `LifeStyle`, `Human`, `Vehicles`), models all perform relatively poorly mainly in *Aesthetic Quality* (shown in Figure 7). This is likely due to the challenges in synthesizing harmonious color schemes, articulated structures, and appealing layouts amidst complex elements. On the other hand, for categories involving complex and intense motions like `Human` and `Vehicle` (see their *Dynamic Degree* in Supplementary File), performance is relatively poor across *all dimen-*

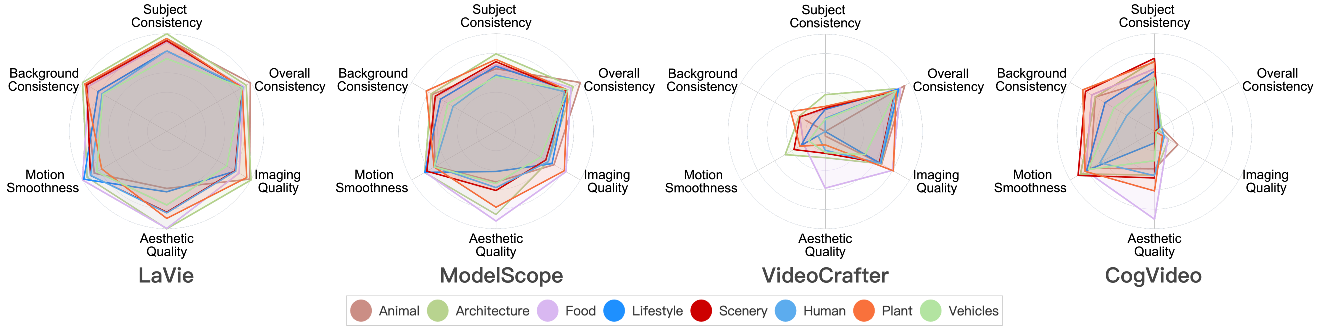


Figure 7. **VBench Results across Eight Content Categories** (best viewed in color). For each chart, we plot the VBench evaluation results across eight different content categories, benchmarked by our *Prompt Suite per Category*. The results are linearly normalized between 0 and 1 for better visibility across categories. See the Supplementary File for comprehensive numerical results, and normalization details.

sions. This suggests that motion complexity and dynamic intensity significantly hinder synthesis, impacting both spatial and temporal dimensions, probably because poor temporal modeling results in distorted and blurred imagery. This highlights the need for improved handling of dynamic motions in video generation models.

· **Challenges of Data Quantity in Handling Complex Categories like Human.** The WebVid-10M dataset [1] allocates 26% of its content to the `Human` category, which is the largest share among the eight categories (see statistics in Supplementary File). However, the `Human` category exhibits one of the poorest results among eight categories (see Figure 7). This suggests that merely increasing data volume may not significantly enhance performance in complex categories like `Human`. A potential approach could involve integrating human-related priors or controls, such as skeletons, to better capture the articulated nature of human appearances and movements.

· **Prioritizing Data Quality Over Quantity in Large-Scale Datasets.** For *Aesthetic Quality*, Figure 7 shows that the `Food` category almost always tends to have the highest scores among all categories. This is corroborated by the WebVid-10M dataset [1], where `Food` ranks highest in *Aesthetic Quality* according to VBench evaluation (refer to Supplementary File for more details), despite comprising just 11% of the total data. This observation suggests that at million scales, data quality might hold greater importance than quantity. Furthermore, *VBench's evaluation dimensions can be potentially useful for cleaning datasets in specified quality dimensions.*

· **Compositionality: T2I versus T2V.** As shown in Figure 6 (a), T2V models significantly underperform in *Multiple Objects* and *Spatial Relationship* compared to T2I models (especially SDXL [71]), which highlights the need to enhance compositionality (*i.e.*, correctly composing multiple objects in the same frame). We believe possible solutions might be: *1)* curating training data incorporating multiple objects with corresponding captions explicitly depicting this compositionality, or *2)* adding intermediate spatial control modules or modalities during video synthesis. Furthermore, the disparity of the text encoders might also account for the performance gap. As T2I models leverage bigger (OpenCLIP ViT-H for SD2.1 [74]) or more sophisticated (CLIP ViT-L & OpenCLIP ViT-G for SDXL [71]) text encoders compared with T2V models (*e.g.*, CLIP ViT-L alone for LaVie), more representative text embeddings could be featuring more accurate object composition comprehension.

6. Conclusion

With the growing focus on video generation, comprehensive evaluation of these models is essential to assess current advancements and guide future research. In this work, we take the first step forward and propose ***VBench***, a comprehensive benchmark suite for evaluating video generation models. With its *multi-dimensional, human-aligned, and insight-rich* properties, VBench could play vital roles for evaluating future video generation models and inspiring further advancements in video generation. We believe that VBench is a significant contribution to the video generation and evaluation community.

Limitations and Future Work. We plan to extend VBench to include more models that recently became available, and extend the evaluations aspects to additional video generation tasks, like image-to-video.

Potential Negative Societal Impacts. We also recognize the importance of considering ethical aspects in future iterations of VBench. While VBench currently does not assess safety and equality dimensions, we urge users to exercise caution with open-sourced video generation models.

Acknowledgement. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s); the National Key R&D Program of China (2022ZD0160201); National Natural Science Foundation of China (62076119, 61921006, 62102150); the Science and Technology Commission of Shanghai Municipality (23YF1461900, 23QD1400800); and Shanghai Artificial Intelligence Laboratory. We thank Shangchen Zhou, Jianyi Wang, and Ruicheng Feng for helpful suggestions.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 6, 8

[2] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *ICCV*, 2023. 3

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[4] Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023. 3

[5] 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 *CVPR*, 2023. 2, 3

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 4

[8] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023. 3

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024. 2

[11] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3

[12] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[13] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023. 3

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 3

[15] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *NeurIPS*, 2022. 2

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[17] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 3

[18] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, 2020. 4

[19] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, 2022. 2

[20] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, and Ziwei Liu. Unitedhuman: Harnessing multi-source data for high-resolution human generation. In *ICCV*, 2023. 2

[21] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 2

[22] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, 2023. 3

[23] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2

[24] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[25] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 3

[26] 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. In *ICLR*, 2024. 3

[27] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 2

[28] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. 3

[29] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-

fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 3, 4, 6

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2, 3

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[32] 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. 3

[33] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[34] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 5, 6

[35] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv: 2307.06350*, 2023. 3, 4

[36] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023. 4

[37] Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*, 2023. 3

[38] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. ReVersion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 3

[39] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-Edit: Fine-grained facial editing via dialog. In *ICCV*, 2021. 2

[40] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM TOG*, 2022. 2

[41] Yuming Jiang, Ziqi Huang, Tianxing Wu, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained 2d and 3d facial editing via dialog. *IEEE TPAMI*, 2023. 2

[42] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2Performer: Text-driven human video generation. In *ICCV*, 2023. 3

[43] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. 3

[44] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 2

[45] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[46] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

[47] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 2

[48] 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

[49] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. *CoRR*, abs/2108.05997, 2021. 4

[50] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3

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

[52] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 2

[53] LAION-AI. aesthetic-predictor. `https://github.com/LAION-AI/aesthetic-predictor`, 2022. 4

[54] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *arXiv preprint arXiv:2311.04287*, 2023. 3

[55] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM MM*, 2019. 2

[56] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 6

[57] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023. 4

[58] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, 2023. 4

[59] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 3

[60] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *NeurIPS*, 2023. 3

[61] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, 2024. 3

[62] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. VideoFusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 2, 3, 6

[63] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3

[64] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 3

[65] Kai Zhang Yujie Lu Xingyu Fu Wenwen Zhuang Wenhu Chen Max Ku, Tianle Li. Imagenhub: Standardizing the evaluation of conditional image generation models. In *ICLR*, 2024. 3

[66] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. *arXiv preprint arXiv:2402.14797*, 2024. 2

[67] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[68] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[69] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*, 2023. 2

[70] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *CVPR*, 2024. 3

[71] 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. 3, 7, 8

[72] 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 preprint arXiv:2303.09535*, 2023. 3

[73] 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 *ICML*, 2021. 2, 3, 4, 6

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 7, 8

[75] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[76] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 2, 3

[77] 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. 2, 3

[78] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3

[79] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2, 3

[80] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[81] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 4

[82] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE TIP*, 2021. 2

[83] 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. 2, 3

[84] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLRW*, 2019. 2

[85] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[86] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2022. 2

[87] 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, 3, 4, 6

[88] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and EditBench: Advancing and evaluating text-guided

image inpainting. *arXiv preprint arXiv:2212.06909*, 2022. 3

[89] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 2

[90] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 3

[91] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 3, 4, 6

[92] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 4

[93] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *ECCV*, 2022. 2

[94] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *arXiv preprint arXiv:2210.05357*, 2022.

[95] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: Temporal distortion-content transformers for video quality assessment. *IEEE TCSVT*, 2023.

[96] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *ICME*, 2023.

[97] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment. *arXiv preprint arXiv:2304.14672*, 2023.

[98] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023.

[99] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable video quality assessment: A database and a language-prompted approach. In *ACM MM*, 2023. 2

[100] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 4

[101] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 2

[102] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 3

[103] 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. 3

[104] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3

[105] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3

[106] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[107] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 2

[108] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[109] 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. 2, 3

[110] Jianfeng Zhang, Hanshu Yan, Zhongcong Xu, Jiashi Feng, and Jun Hao Liew. Magicavatar: Multimodal avatar generation and animation. *arXiv preprint arXiv:2308.14748*, 2023. 3

[111] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[112] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 5

[113] 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. 3