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/

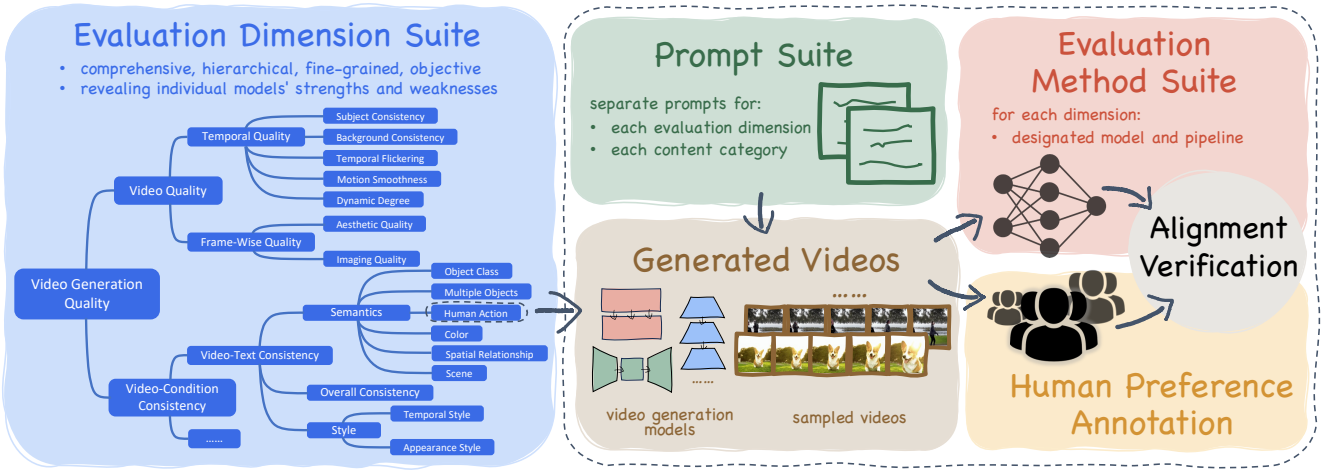


Figure 1. **Overview of VBench.** We propose VBench, a comprehensive benchmark suite for video generative models. We design a comprehensive and hierarchical **Evaluation Dimension Suite** to decompose "video generation quality" into multiple well-defined dimensions to facilitate fine-grained and objective 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

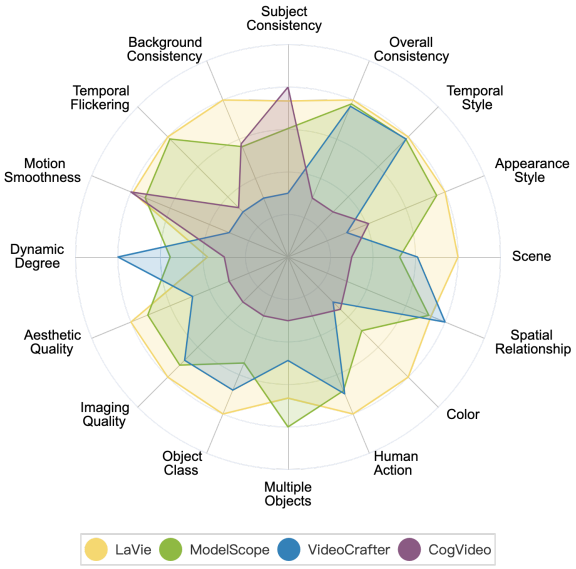


Figure 2. **VBench Evaluation Results of Video Generative Models.** We visualize the evaluation results of four video generation models in 16 VBench dimensions. We normalize the results per dimension for clearer comparisons. For comprehensive numerical results, please refer to Table 1.

1. Introduction

Image generation models have made rapid progress in the past few years, such as Variational Autoencoders (VAEs) [59], Generative Adversarial Networks (GANs) [9, 25, 26, 30, 48, 52–55, 74], vector quantized (VQ) based approaches [22, 49, 97], and diffusion models [38, 89, 91]. This fuels recent explorations in video generation [8, 35, 41, 72, 88, 98, 104, 116, 124], 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) [87], Fréchet inception distance (FID) [37], Fréchet Video Distance (FVD) [95, 96], and CLIPSIM [83] are inconsistent with human judgement [20, 79]. Meanwhile, the Video Quality Assessment (VQA) methods [63, 94, 108–114] 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 [19, 38, 89, 91] have achieved significant progress in image synthesis [31, 46, 47, 78, 81, 84, 86], and enabled a line of works towards video generation [8, 13, 28, 33–35, 39, 40, 50, 58, 72, 88, 98, 104, 118, 124, 132, 133]. Many recent diffusion-based works [35, 72, 98, 104] are text-to-video (T2V) models. Other guidance modalities are also available, including image-to-video [14, 16, 23, 122], video-to-video [12, 67, 80, 82, 121], and a variety of control maps [15, 51, 58, 73, 103, 125, 126] 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) [87], Fréchet inception distance (FID) [37], Fréchet Video Distance (FVD) [95], and CLIPSIM [83] for evaluation. The UCF-101 [92] dataset's class labels often serve as text prompts for IS, FID, and FVD, whereas MSR-VTT [120]'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 [6, 7, 44, 61, 86, 99] are proposed to assess various capabilities like compositionality [44] and editing ability [7, 99]. However, video generative models still lack comprehensive evaluation benchmarks for detailed and human-aligned feedback. Our work differs from concurrent research [70, 71] 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 [95] 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.

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 different evaluation dimensions, where each focusing on a different aspect of temporal quality. We briefly introduce each dimension here. *Please refer to the Supplementary File for the detailed definition and evaluation method of each dimension.*

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 [10] feature similarity across frames.

Temporal Quality - Background Consistency. We evaluate the temporal consistency of the background scenes by calculating CLIP [83] 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. Both *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 [66] to evaluate the smoothness of generated motions (see the detailed method in Supplementary File).

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 [93] 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 [60]. 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 [57] image quality predictor trained on the SPAQ [24] 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 [115] 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 [65] 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 [115] to provide color captioning, and compare against the expected color.

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 [44].

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 [45] 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 [83] 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 [105] 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 [105] on general text prompts as an aiding metric to reflect both semantics and style consistency.

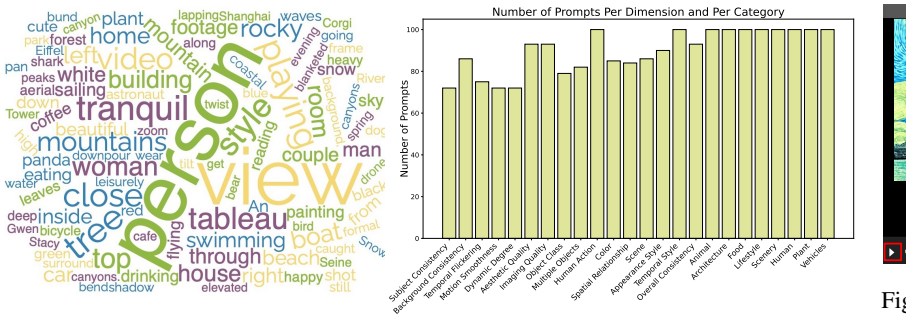


Figure 3. **Prompt Suite Statistics.** The two graphs provide an overview of our prompt suites. ***Left:*** the word cloud to visualize word distribution of our prompt suites. ***Right:*** the number of prompts across different evaluation dimensions and different content categories.

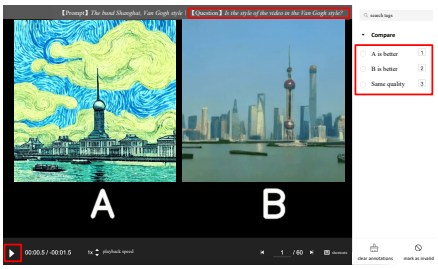


Figure 4. **Interface for Human Preference Annotation**. ***Top:*** prompt and question. ***Right:*** choices that annotators can make. ***Bottom left:*** control for stop and playback.

For each dimension, please refer to the Supplementary File for: 1) details of its definition, 2) positive and negative examples (*i.e.*, synthesized videos) of each dimension, and 3) detailed evaluation method and pipeline implementations.

3.2. Prompt Suite

The sampling procedure of current diffusion-based video generation models [35, 98, 104] is time-consuming (*e.g.*, 90 seconds per video for LaVie [104], and more than 2 minutes per video for CogVideo [41]). 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 set of around 100 prompts as the test cases. The prompt suite is carefully curated to test the specific ability corresponding 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. For the dimension *"Object Class"*, 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 [56], 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 each of the 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) [130] (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*

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 [104]	91.41%	**97.47%**	**98.30%**	96.38%	49.72%	**54.94%**	**61.90%**	**91.82%**
ModelScope [72, 98]	89.87%	95.29%	98.28%	95.79%	66.39%	52.06%	58.57%	82.25%
VideoCrafter [35]	86.24%	92.88%	97.60%	91.79%	**89.72%**	44.41%	57.22%	87.34%
CogVideo [41]	**92.19%**	95.42%	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 [104]	33.32%	**96.80%**	**86.39%**	34.09%	**52.69%**	**23.56%**	**25.93%**	**26.41%**
ModelScope [72, 98]	**38.98%**	92.40%	81.72%	33.68%	39.26%	23.39%	25.37%	25.67%
VideoCrafter [35]	25.93%	93.00%	78.84%	**36.74%**	43.36%	21.57%	25.42%	25.21%
CogVideo [41]	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%

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 (see more details in the Supplementary File).

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 form the instruction data, which is then used to fine-tune the pre-trained VideoChat [64] 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 [104], ModelScope [72, 98], VideoCrafter [35], and CogVideo [41] 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 [5] dataset quality in each VBench dimension.

Empirical Max. For most dimensions, to approximate the maximum achievable values, we first retrieve WebVid-10M [5] videos according to our prompt suites. We use CLIP [83] 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 [5] 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 calcu-

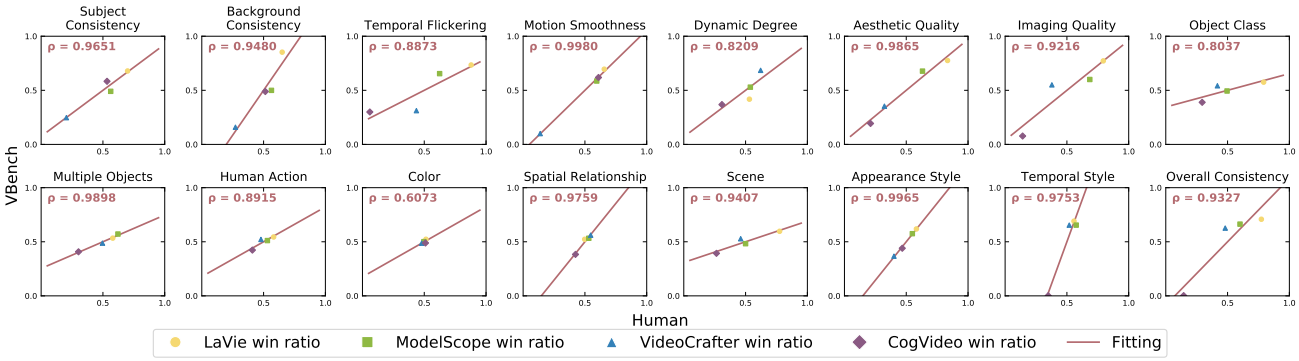
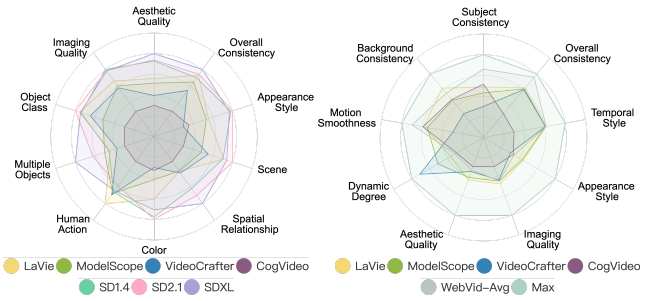


Figure 5. **Validate VBench's Human Alignment.** Our experiments show that ***VBench evaluations across all dimensions closely match human perceptions.*** Each plot shows the alignment verification result of a specific VBench dimension. In each plot, a dot represents the human preference win ratio (horizontal 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.

lated as the total score divided by the total number of pairs-wise 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 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 [84], SD2.1 [84], and SDXL [81]. 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 dif-

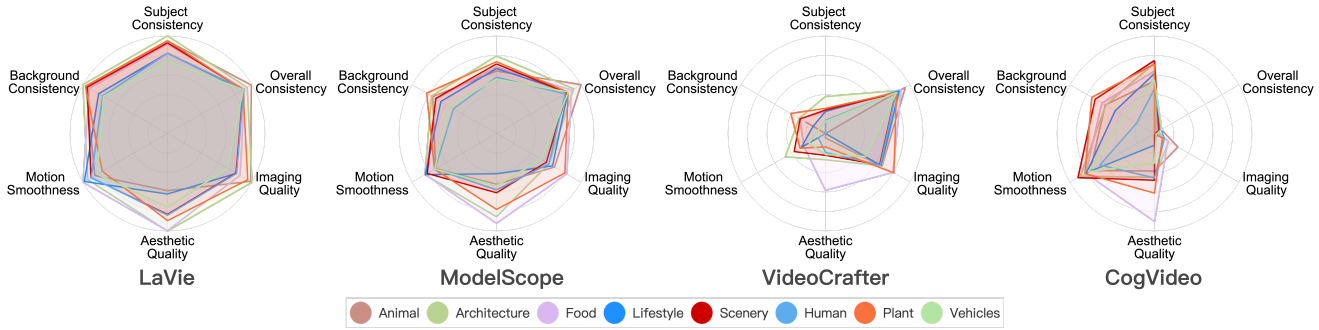


Figure 7. **VBench Results across Eight Content Categories** (best viewed in color). For each chart, we plot the VBench evaluation results across eight different content categories, benchmarked by our *Prompt Suite per Category*. The results are linearly normalized between 0 and 1 for better visibility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

ferent 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 dimensions*. 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 [5] 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 [5], 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 [81]), 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 [84]) or more sophisticated (CLIP ViT-L & OpenCLIP ViT-G for SDXL [81]) 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 expand VBench to include more models when they become available and extend the evaluations 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. We would like to thank Shangchen Zhou, Jianyi Wang, and Ruicheng Feng for their helpful suggestions.

References

[1] Gen-2. Accessed September 25, 2023 [Online] https://research.runwayml.com/gen2, 2023. 27

[2] Morph studio. Accessed September 25, 2023 [Online] https://www.morphstudio.com/, 2023.

[3] Pika labs. Accessed September 25, 2023 [Online] https://www.pika.art/, 2023.

[4] Zeroscope-xl. Accessed September 25, 2023 [Online] https://huggingface.co/cerspense/zeroscope_v2_XL, 2023. 27

[5] 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, 25, 26

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

[7] 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

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2, 3, 27

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

[10] 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, 14

[11] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023. 27

[12] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023. 3, 27

[13] 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

[14] 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, 27

[15] 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, 2023. 3, 27

[16] 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, 27

[17] Ernie Chu, Shuo-Yen Lin, and Jun-Cheng Chen. Video controlnet: Towards temporally consistent synthetic-to-real video translation using conditional image diffusion models, 2023. 27

[18] Paul Couairon, Clément Rambour, Jean-Emmanuel Haugeard, and Nicolas Thome. Videdit: Zero-shot and spatially aware text-driven video editing. *arXiv preprint arXiv:2306.08707*, 2023. 27

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

[20] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *NeurIPS*, 2022. 2, 25

[21] Patrick Esser, Robin Rombach, and Björn Ommer. A note on data biases in generative models. In *NeurIPS Workshop*, 2020. 27

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

[23] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 3, 27

[24] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, 2020. 4, 17

[25] 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

[26] 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

[27] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation, 2023. 27

[28] 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

[29] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 27

[30] 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

[31] 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

[32] Xianfan Gu, Chuan Wen, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 27

[33] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3, 27

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

[35] 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, 5, 6, 15, 24, 25, 27

[36] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 27

[37] 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

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

[39] 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

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

[41] 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, 15, 24, 25, 27

[42] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 27

[43] Jiahui Huang, Leonid Sigal, Kwang Moo Yi, Oliver Wang, and Joon-Young Lee. Inve: Interactive neural video editing, 2023. 27

[44] 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, 18

[45] 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, 18

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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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, 27

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

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

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

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

[56] 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, 18, 20

[57] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. *CoRR*, abs/2108.05997, 2021. 4, 16

[58] 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, 27

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

[60] LAION-AI. aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022. 4, 16

[61] 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

[62] Yao-Chih Lee, Ji-Ze Genevieve Jang Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing demo. *arXiv preprint arXiv:2301.13173*, 2023. 27

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

[64] 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, 22

[65] 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, 18

[66] 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, 16

[67] 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, 27

[68] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 19, 20

[69] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control, 2023. 27

[70] 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. *arXiv preprint arXiv:2310.11440*, 2023. 3

[71] 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

[72] 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, 15, 24, 25, 27

[73] 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, 27

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

[75] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. WALDO: Future video synthesis using object layer decomposition and parametric flow prediction. In *ICCV*, 2023. 27

[76] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 27

[77] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023. 27

[78] 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

[79] 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

[80] 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. *arXiv preprint arXiv:2308.07926*, 2023. 3, 27

[81] 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, 16, 27

[82] 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, 27

[83] 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, 14, 19

[84] 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, 25, 27

[85] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine

tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 14

[86] 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

[87] 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

[88] 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, 27

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

[90] Xue Song, Jingjing Chen, Bin Zhu, and Yu-Gang Jiang. Text-driven video prediction. *arXiv preprint arXiv:2210.02872*, 2022. 27

[91] 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

[92] 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

[93] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 4, 15, 16

[94] 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*, 30:4449–4464, 2021. 2

[95] 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

[96] 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

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

[98] 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, 5, 6, 15, 24, 25, 27

[99] 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

[100] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 27

[101] Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 27

[102] Xiaodong Wang, Chenfei Wu, Shengming Yin, Minheng Ni, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Fan Yang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Learning 3d photography videos via self-supervised diffusion on single images, 2023. 27

[103] 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, 27

[104] 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, 5, 6, 15, 24, 25, 27

[105] 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, 19

[106] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv preprint arXiv:2305.03989*, 2023. 27

[107] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 27

[108] 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

[109] 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.

[110] 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 Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

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

[112] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Wenxiu Sun 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.

[113] 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 *ICCV*, 2023.

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

[115] 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, 17

[116] 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, 27

[117] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 27

[118] 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

[119] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 27

[120] 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

[121] 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, 27

[122] 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, 2023. 3, 27

[123] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer, 2023. 27

[124] 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, 27

[125] Jianfeng Zhang, Hanshu Yan, Zhongcong Xu, Jiashi Feng, and Jun Hao Liew. Magicavatar: Multi-modal avatar generation and animation. In *arXiv*, 2023. 3, 27

[126] 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, 27

[127] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models, 2023. 27

[128] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023.

[129] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023. 27

[130] 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, 20

[131] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014. 19

[132] 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

[133] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. 3

VBench: Comprehensive Benchmark Suite for Video Generative Models

Supplementary Material

In this ***supplementary file***, we provide more details on *Evaluation Dimension Suite* and *Evaluation Method Suite* in Section G, and elaborate on *Prompt Suite* details in Section H. We then provide further explanations on *Human Preference Annotations* in Section I, and more implementation details on our experiments and visualizations in Section J. The potential societal impacts of our work are discussed in Section K. We also discuss our limitations in Section L. Finally, in Section M, we provide additional experimental results used to support the visualizations and insights in the main paper.

A ***demo video*** is also provided along with this supplementary file to illustrate VBench and show video examples of each dimension.

G. More Details on Evaluation Dimension and Method Suite

G.1. Video Quality

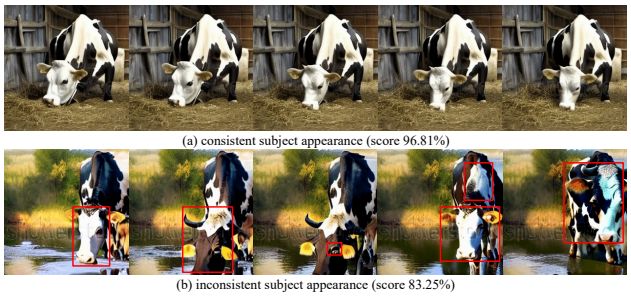


Figure A8. **Visualization of *Subject Consistency***. We demonstrate different degrees of subject consistency, as indicated by our *Subject Consistency* score (the larger the better) **(a)** The cow has a relatively consistent look throughout across different frames. **(b)** The cow shows inconsistency in its appearance over time. The red boxes indicate areas of subject inconsistency.

Subject Consistency. When there is a subject (*e.g.*, a cow, a person, a car, or a cat) in the video, it is important that the subject looks consistent throughout the video (*i.e.*, whether it is still the same thing or the same person). For example, in Figure A8, the cow in the top row remains consistent across different frames, while the cow in the bottom row shows changes in appearance between frames. To evaluate subject consistency, we employ DINO [10] to extract features from each frame to represent the subject. Since DINO is not trained to disregard the differences within subjects of the same class [85], its feature extraction is particularly sensitive to the identity variations of the subject within the video, thereby making it a suitable tool for evaluating subject consistency. Specifically, for each video, the subject consistency score is calculated as:

$$S_{subject} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}(\langle d_1 \cdot d_t\rangle + \langle d_{t-1} \cdot d_t\rangle), \quad (1)$$

where d_i is the DINO image feature of the i^{th} frame, normalized to unit length, and $\langle\cdot\rangle$ is the dot product operation for calculating cosine similarity. For each frame, we calculate the cosine similarity with the first frame and its preceding frame, take the average, and then compute the mean over all the non-starting video frames. We average the score $S_{subject}$ for all the videos generated by one model as the final score of the model.

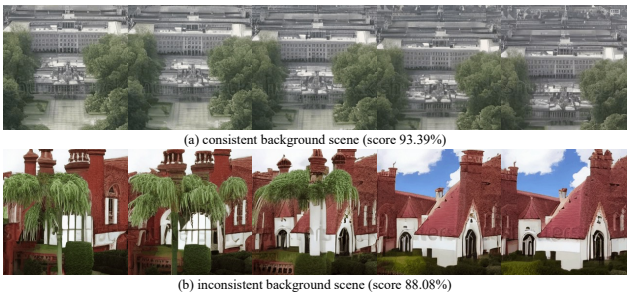
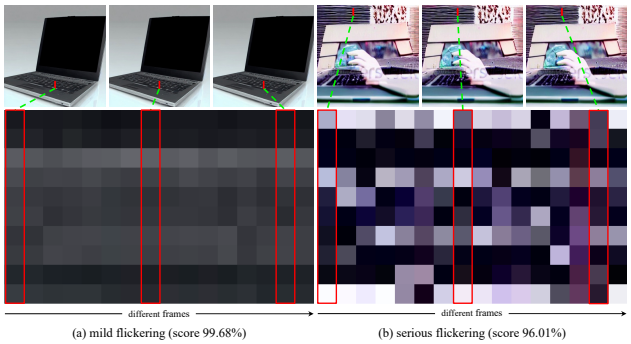


Figure A9. **Visualization of *Background Consistency***. We showcase varying levels of background consistency, as indicated by our *Background Consistency* metrics (larger values denote better consistency) **(a)** The background scene maintains a high degree of consistency (*i.e.*, still the same scene) across different frames. **(b)** The background exhibits noticeable distortion and abrupt changes over time.

Background Consistency. Beyond the focus on the foreground subject, maintaining a consistent background scene across different frames is equally important. For example, in Figure A9, in the top row, the scene maintains a consistent appearance as the camera moves, while in the bottom row, the entire scene undergoes significant changes within a few frames. For each video frame, we employ the CLIP [83] image encoder to extract its feature vector. We then compute the background consistency metric, which is similar to the method used for *Subject Consistency*:

$$S_{background} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}(\langle c_1 \cdot c_t\rangle + \langle c_{t-1} \cdot c_t\rangle), \quad (2)$$

where c_i represents the CLIP image feature of the i^{th} frame, normalized to unit length.



(a) mild flickering (score 99.68%) (b) serious flickering (score 96.01%)

Figure A10. **Visualization of *Temporal Flickering*.** We demonstrate different degrees of temporal flickering, with a mild occurrence in (a), and a severe occurrence in (b), both reflected by our flicker score metrics (the larger the better). To visualize temporal flickering, given a generated video (***top row***), we extract a small segment of pixels (marked as the red segment) from each frame at the same location and stack them in frame order (***bottom row***). **(a)** Pixel values do not vary abruptly, and the video suffers less from flickering. **(b)** Pixel values vary abruptly and frequently across different frames, showing strong temporal flickering. Our evaluation metrics also give a lower score.

Table A2. ***Dynamic Degree* on Three Benchmarks.** We report the *Dynamic Degree* metrics on three *Temporal Flickering* benchmarks. We use videos from the Subject Consistency dimension as the "Dynamic Benchmark", videos from the Background Consistency dimension as the "Semi-Dynamic Benchmark", and videos from the temporal flickering dimension as the "Static Benchmark".

Models	Static Benchmark	Semi-Dynamic Benchmark	Dynamic Benchmark
LaVie [104]	0.00%	6.51%	49.72%
ModelScope [72, 98]	0.00%	33.72%	66.39%
VideoCrafter [35]	0.00%	51.63%	89.72%
CogVideo [41]	0.00%	14.19%	42.22%

Temporal Flickering. For real videos, temporal flickering is usually a result of frequent lighting variation, or shaky camera motions during the video capture process. However, for generated videos, temporal flickering is an intrinsic property of the video generation model, usually caused by imperfect temporal consistency at local and high-frequency details. In generated videos, temporal inconsistency can be attributed to various types of issues, including temporal flickering, unnatural motions, subject inconsistency *etc*. To disentangle the evaluation of temporal flickering from other aspects, we use static video scenes (*i.e*., no apparent motions) as the test cases (We use carefully designed prompts to generate static scenes for video sampling. To further ensure that the evaluation is conducted on static videos without apparent motions, we employ an optical flow estimator [93] to filter out videos and only keep the static videos). We calculate the frame-by-frame temporal flickering degree

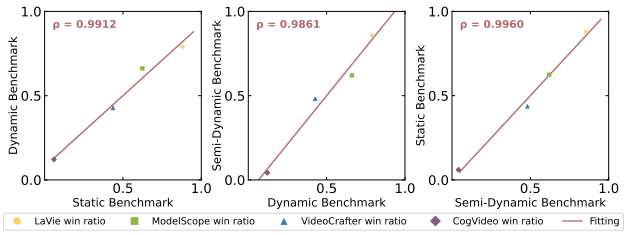


Figure A11. **Temporal Flickering Human Preference across Different Dynamic Degrees.** In each plot, a dot represents the human preference win ratio, where the horizontal and vertical axes correspond to two different benchmarks with different dynamic degrees. We linearly fit a straight line to visualize the correlation and calculate the correlation (ρ) for each dimension. We observe that the human preferences in terms of temporal flickering on these three benchmarks have high mutual correlations of around 99%.

with the following formula:

$$S_{flicker} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T-1}\sum_{t=1}^{T-1} MAE(f_i^t, f_i^{t+1})\right), \quad (3)$$

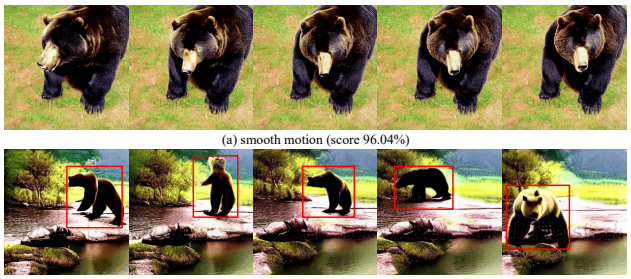
where N is the number of videos generated by a model, T is the number of frames per video, f_i^t is the frame t in video i, and MAE is the Mean Absolute Error between two consecutive frames over all pixel locations. We then normalize the temporal flickering degree to $[0, 1]$ as follows:

$$S_{flicker-norm} = \frac{255 - S_{flicker}}{255}, \quad (4)$$

where a higher score implies less flickering, and thus better video perceptual quality in terms of temporal flickering.

To verify that the strength of motions (*i.e*., large motion or small motion) in videos does not significantly impact the model's ranking in terms of temporal flickering, we conduct separate human evaluations for the level of temporal flickering on videos with different dynamic degrees, and show in Figure A11 that model ranking in terms of temporal flickering does not vary based on the dynamic degree of test videos. For videos of high dynamic degrees, we use videos from the *Subject Consistency* dimension's prompt suite, and term as the "Dynamic Benchmark". For videos that exhibit lower dynamic degrees but remain non-static, we use videos sampled from the *Background Consistency* dimension's prompt suite, and label them as the "Semi-Dynamic Benchmark". Additionally, the "Static Benchmark" refers to the videos sampled from the prompt suite for the *Temporal Flickering* dimension. We show the dynamic degree of videos in these three benchmarks in Table A2. In Figure A11, we show that the human win ratio in terms of temporal flickering on three benchmarks is almost perfectly correlated with each other, with a correlation of around 99% between any two benchmarks. Therefore, we believe the

degree of motion is disentangled with the temporal flickering ranking in video generative models, and we use the "Static Benchmark" for easier and more focused evaluation on *Temporal Flickering*.



(a) smooth motion (score 96.04%)

(b) unnatural motion (score 88.47%)

Figure A12. **Visualization of *Motion Smoothness***. We investigate various levels of motion smoothness, ranging from being smooth as depicted in (a) to highly erratic as depicted in (b), as indicated by our motion score metrics (larger values denote better smoothness). The red boxes indicate areas of discontinuous motion.

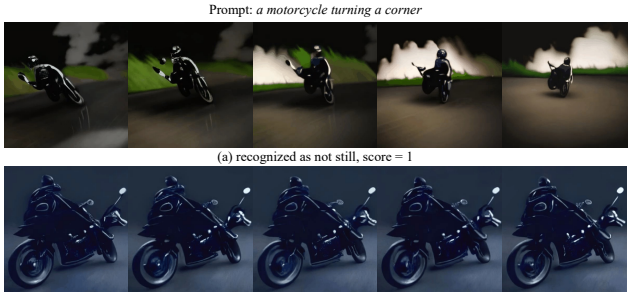
Motion Smoothness. To evaluate whether the motion in the generated video is smooth and follows the physical law of the real world, we make use of the frame-by-frame motion prior to video frame interpolation models. Specifically, video frame interpolation models usually assume real-world motions within a very short time period (*i.e.*, a few consecutive frames) to be linear or quadratic and synthesize the non-existing intermediate frames based on this assumption. Given a generated video consisting of frames $[f_0, f_1, f_2, f_3, f_4 ..., f_{2n-2}, f_{2n-1}, f_{2n}]$, we manually drop the odd-number frames to obtain a lower-frame-rate video $[f_0, f_2, f_4 ..., f_{2n-2}, f_{2n}]$, and use video frame interpolation [66] to infer the dropped frames $[\hat{f}_1, \hat{f}_3, ..., \hat{f}_{2n-1}]$. We then compute the Mean Absolute Error (MAE) between the reconstructed frames and the original dropped frames. The calculated MAE is normalized in the same way as Equation 4, so that the final score falls into $[0, 1]$, with a larger number implying smoother motion.

Dynamic Degree. Based on our observations, some models tend to generate static videos even when the prompt includes descriptions of movement. This results in a noticeable advantage for these models in evaluations of other temporal consistency dimensions, leading to unfair comparisons. This dimension is designed to assess the extent to which models tend to generate non-static videos. We use RAFT [93] to estimate optical flow strengths between consecutive frames of a generated video. We then take the average of the largest 5% optical flows (considering the movement of small objects in the video) as the basis to determine whether the video is static. The final dynamic degree score is calculated by measuring the proportion of non-static videos generated by the model.

Aesthetic Quality. Aesthetic Quality takes photographic layout rules, the richness and harmonies of colors, the artistic quality of the subjects, etc into account. We adopt an image aesthetic quality predictor to evaluate the generated videos frame by frame. We use the LAION aesthetic predictor [60] to give a 0-10 rating for each frame, linearly normalize the score to 0-1, and calculate the average score of all synthetic frames as the final video aesthetic score.

Imaging Quality. Imaging quality mainly considers the low-level distortions presented in the generated video frames *(e.g., over-exposure, noise, blur)*. We use the MUSIQ [57] image quality predictor trained on the

Prompt: *a motorcycle turning a corner*

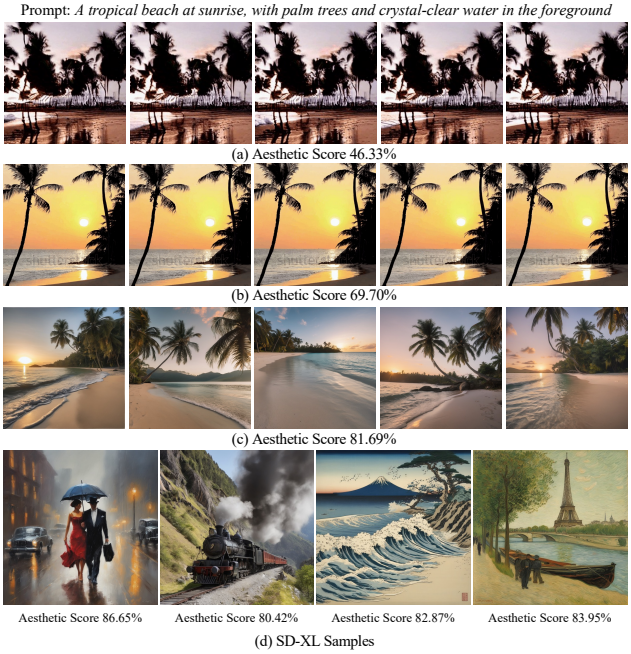


(a) recognized as not still, score = 1

(b) recognized as still, score = 0

Figure A13. **Visualization of *Dynamic Degree***. We present generated examples of different degrees of motion. **(a)** In the video, there is obvious motion of the camera and the object, which is identified as dynamic. **(b)** The video remains almost unchanged from the start to the end and is identified as static.

Prompt: *A tropical beach at sunrise, with palm trees and crystal-clear water in the foreground*



(a) Aesthetic Score 46.33%

(b) Aesthetic Score 69.70%

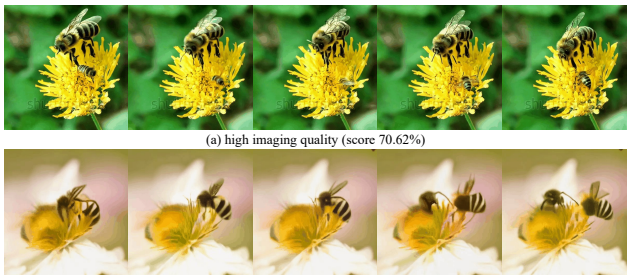
(c) Aesthetic Score 81.69%

Aesthetic Score 86.65% Aesthetic Score 80.42% Aesthetic Score 82.87% Aesthetic Score 83.95%

(d) SD-XL Samples

Figure A14. **Visualization of *Aesthetic Quality***. We demonstrate video frames with varying degrees of aesthetic quality in (a), (b), and (c), which are effectively reflected by our aesthetic score metrics (higher indicating better). In (d), we showcase images with high aesthetic scores sampled from SDXL [81].

SPAQ [24] dataset, which is capable of handling variable-sized aspect ratios and resolutions. The frame-wise score is linearly normalized to $[0, 1]$ by dividing 100, and the final score is then calculated by averaging the frame-wise scores across the entire video sequence.

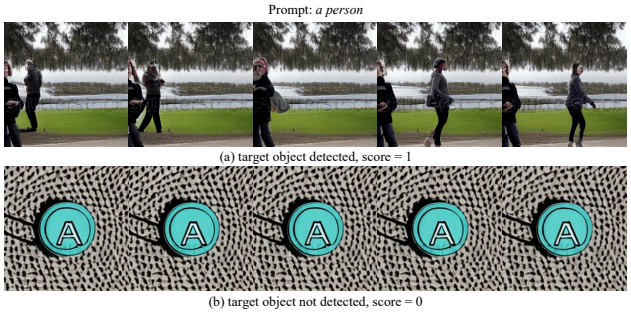


(a) high imaging quality (score 70.62%)

(b) low imaging quality (score 31.27%)

Figure A15. **Visualization of *Imaging Quality***. We present examples of generated videos with high imaging quality scores in (a), and low imaging quality scores (where the video is blurry and over-exposed) in (b).

G.2. Video-Condition Consistency

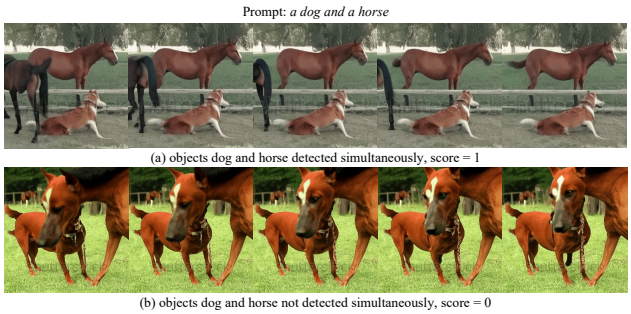


(a) target object detected, score = 1

(b) target object not detected, score = 0

Figure A16. **Visualization of *Object Class***. We demonstrate generation examples for the target object at varying degrees, as reflected by the object score metrics (where 1 represents success, and 0 represents failure). **(a)** The target object "person" is successfully generated in the video. **(b)** The synthesized video does not contain the target object.

Object Class. When a user specifies a certain type of object in the text prompt, we aim to evaluate whether the model can generate an object of the specified type. To this end, we use GRiT [115] to detect objects in each frame of the generated video and check whether the specified object class is successfully detected in these frames. Subsequently, we report the proportion of frames in which the corresponding object class has been successfully detected. We employ GRiT for this dimension, as well as several other semantics dimensions such as *Multiple Objects*, *Color*, and *Spatial Relationship* for two reasons: *1)* GRiT is a versatile framework that can handle both detection and captioning tasks, predicting diverse object attributes, so that the VBench can use the same framework across different dimensions and save users from installing multiple frameworks or downloading multiple pre-trained models. *2)* GRiT demonstrates reliable performance in evaluating our designated dimensions, with comparable performance with the state-of-the-art object detectors [115], and good alignment with human perception in terms of "correct detection" as validated by the human preference results in main paper Figure 5.

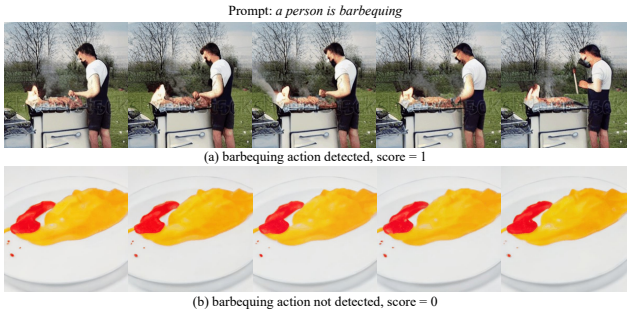


(a) objects dog and horse detected simultaneously, score = 1

(b) objects dog and horse not detected simultaneously, score = 0

Figure A17. **Visualization of *Multiple Objects***. We showcase instances of generating multiple objects within a video simultaneously at different levels, as indicated by our multiple objects score metrics (where 1 signifies success, and 0 denotes failure). **(a)** The video effectively generates multiple required objects (*i.e.*, dog and horse). **(b)** The video fails to produce the dog and horse at the same time.

Multiple Objects. Other than generating a single object, compositionality is also an essential aspect of video generation. Suppose the user requires generating multiple objects, we use GRiT for frame-wise object detection. For each frame, we check whether all the user-requested objects simultaneously appear in each frame. We then report the proportion of frames in which all the required objects have been successfully detected.



(a) barbequing action detected, score = 1

(b) barbequing action not detected, score = 0

Figure A18. **Visualization of *Human Action***. We showcase examples of generating the target action at different levels, as indicated by our action score metrics (where 1 denotes success, and 0 denotes failure). **(a)** The video successfully generates the barbequing action. **(b)** The video does not generate the target action.

Human Action. In the process of video synthesis from textual prompts, both the mentioned subjects in the prompt and the corresponding actions they engage in are important.

Given the remarkable emergence of high-quality human-centric generated videos, we believe it is necessary to ensure that human subjects depicted in videos accurately execute the specific actions described by the textual prompts. To this end, we use the Kinetics-400 dataset [56] as a reference due to its comprehensive characterization of diverse human actions. To evaluate the accuracy of the generated videos, we uniformly sample 16 frames from each video and apply UMT [65], which achieves the state-of-the-art classification performance on the Kinetics-400 dataset among open-sourced models to classify the action. The top 5 results with logits bigger than 0.85 are preserved as ground-truth candidates, and we check whether the actions mentioned in the text prompt appear in the ground-truth candidates. The average percentage of all classification results is reported to assess whether the generated videos have human actions aligned with the text prompts.

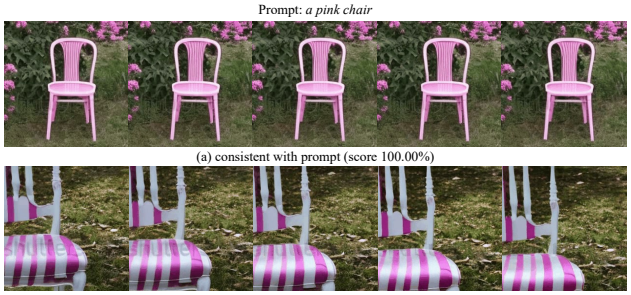


Figure A19. **Visualization of *Color***. We present examples of generating the target color within videos, depicting various levels of success through our color score metrics (larger denotes better). **(a)** The video accurately generates the target color. **(b)** The video only generated the target color in certain parts.

Color. To evaluate whether the color of an object is consistent with the specified condition, we use GRiT's captioning ability to describe colors, with slight modification to the GRiT pipeline. To remove the influence of the *Object Class* dimension's ability, we only consider videos where the object has been successfully generated. Specifically, GRiT identifies the bounding boxes of objects, which are then fed to two text decoders: one for predicting categories and the other for generating dense captions on the synthesized video frame. We then verify if the corresponding object's color is successfully captioned in all frames. Among the frames where the corresponding object is generated and the caption contains color information, we compute the percentage of frames where the color required by the text prompt is successfully captioned.

Spatial Relationship. We focus on *left-right* and *top-bottom* relationships and evaluate whether the video content adheres to the spatial relationship specified by the text prompts. Inspired by the T2I-CompBench [44] evaluation, we compute the spatial relationship accuracy based on the

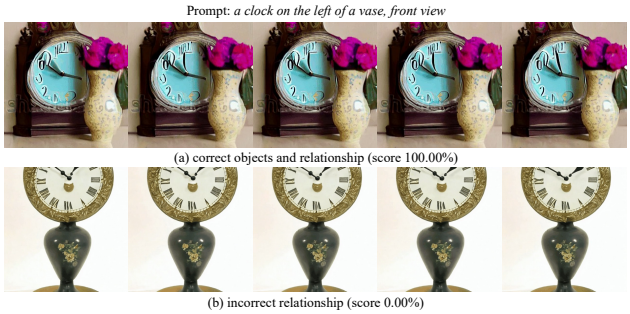


Figure A20. **Visualization of *Spatial Relationship***. We show examples of generating the spatial relationships mentioned in the prompt within videos. **(a)** The video successfully captures the spatial relationship and objects described in the prompt. **(b)** The generated video does not contain the intended relationship.

horizontal and vertical positioning of object pairs. During evaluation, distances on the designated axis (e.g., left-right) are expected to be greater than those on the other orientation (e.g., top-bottom). Under this condition, we observe the intersection over the union metric (IoU) of two objects to obtain the final score, where IoU values that fall below a specified threshold result in a score of 100%, and the values exceeding the threshold are multiplied by a coefficient based on the IoU to determine the final score. We use GRiT to detect the objects and their locations within the generated video frames, and we also calculate the Intersection over Union (IoU) of the two objects' bounding boxes as the final spatial relationship score coefficient.

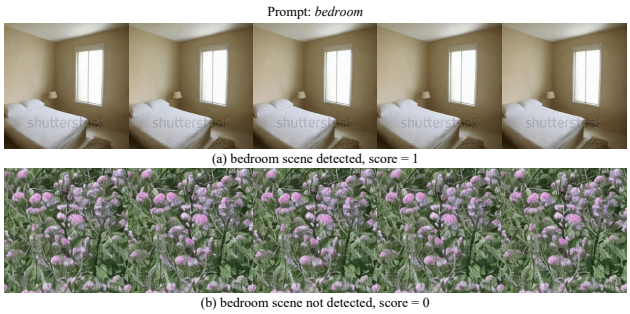
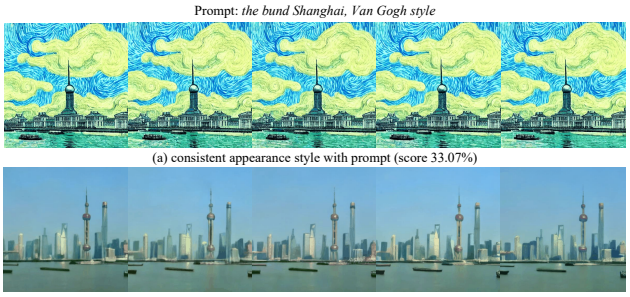


Figure A21. **Visualization of *Scene***. We present examples of generating the required scene (where 1 represents success, and 0 indicates failure). **(a)** The required scene is generated successfully. **(b)** The video does not show the scene as required.

Scene. For a scenario described by the text prompt, we need to evaluate whether the synthesized video is consistent with the intended scene. For example, when prompted to "ocean", the generated video should be "ocean" instead of "river". We use Tag2Text [45] to caption the generated scenes, and then check the correspondence with scene descriptions in the text prompt. Specifically, each word related to the scene in the text prompt needs to appear in the predicted caption, but the word order can be different. We then

report the proportion of frames in which the corresponding scene has been successfully generated.

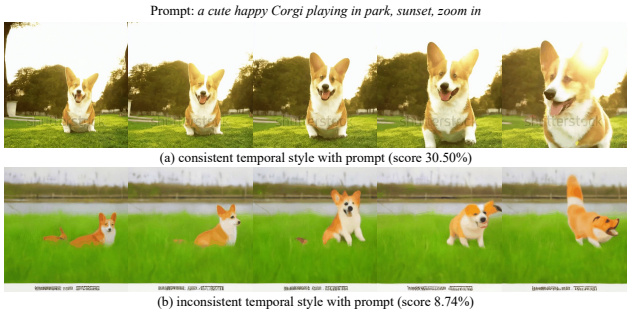


(a) consistent appearance style with prompt (score 33.07%)

(b) inconsistent appearance style with prompt (score 20.93%)

Figure A22. **Visualization of *Appearance Style***. We demonstrate examples of generating the required appearance style within videos, showcasing different levels of success as assessed by our appearance style score metrics. **(a)** The generated video follows the requested Van Gogh style. **(b)** The video does not show the desired appearance style.

Appearance Style. For stylized video generation, we first extract the style description in the text prompt, then evaluate the video-text feature similarity to assess appearance style consistency. Specifically, We use CLIP [83] to extract features from each frame and the text, and then compute the mean cosine similarity of the normalized features. CLIP demonstrates robust zero-shot performance in perceiving textual descriptions of styles, aiding our evaluation of style consistency.

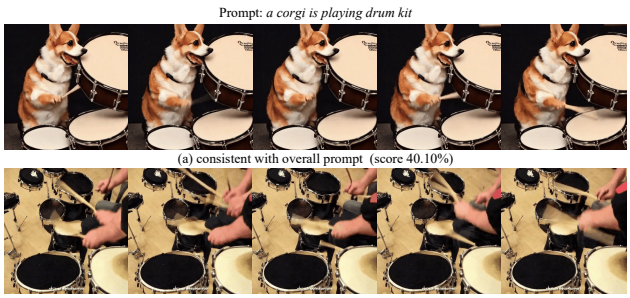


(a) consistent temporal style with prompt (score 30.50%)

(b) inconsistent temporal style with prompt (score 8.74%)

Figure A23. **Visualization of *Temporal Style***. We demonstrate two different generated videos to show the consistency of their temporal style with the prompt at various degrees, measured by our temporal style score. **(a)** The generated video follows the "zoom in" temporal style. **(b)** The video's temporal style does not align with the prompt.

Temporal Style. In videos, style is not only spatially narrated in individual frames, but also temporally revealed in different types of object motions and camera motions. For example, we are interested in whether the text prompt specifies "zoom in" or "zoom out", "pan left" or "pan right", and whether the generated video can show such kind of camera motion. Additionally, there are different types of other temporal styles like "super slow motion", "camera shaking", and "racking focus". In terms of temporal awareness, ViCLIP [105] is pre-trained on a diverse 10M video-text dataset, which shows strong zero-shot learning capabilities in video-text retrieval tasks. When a video is generated based on a specified temporal style, we use ViCLIP to calculate the video-text feature similarity to reflect temporal style consistency.



(a) consistent with overall prompt (score 40.10%)

(b) inconsistent with overall prompt (score 28.87%)

Figure A24. **Visualization of *Overall Consistency***. We demonstrate different examples that illustrate the extent to which they align with the prompt, as measured by our overall score metrics (larger values denote better consistency). **(a)** The video aligns closely with the prompt. **(b)** The video lacks alignment with the target concept.

Overall Consistency. We also use overall video-text consistency computed by ViCLIP as an aiding metric to reflect both semantics and style consistency, where the text prompts contain different semantics and styles.

H. More Details on Prompt Suite

H.1. Prompt Suite per Evaluation Dimension

For each VBench dimension, we carefully designed around 100 prompts as the test cases. *For semantics-related prompt suites, we provide clear semantics labels to each prompt in the prompt suites to facilitate efficient and accurate evaluation.* For example, we provide the object class labels for prompt suites of *Object Class*, *Multiple Objects*, and *Spatial Relationship*. We also provide color labels for *Color* prompts, relationship tags for *Spatial Relationship* prompts, and style labels for *Appearance Style*. We detail the prompt suite for each dimension as follows.

Subject Consistency. We choose 19 representative living or movable object categories from the COCO [68] dataset's 80 object categories. These categories encompass animals and transportation-related items. Each object category is associated with a set of carefully crafted actions or movements, ensuring logical coherence between the actions and their respective objects. A list of distinct prompts used for evaluating subject consistency is therefore created.

Background Consistency. We carefully select a list of distinct and representative scenes from the Places365 [131] dataset, aiming to include a diverse set of scenes within a

limited number of prompts. The selected scenes contain indoor, modern, rural, and various other settings, thereby ensuring the representation of a wide range of environmental contexts. This prompt suite is applied to both the *Background Consistency* dimension and the *Scene* dimension.

Temporal Flickering. To more effectively evaluate temporal flickering, it is essential to eliminate interference from other temporal dimensions. According to observations in Section G, whether the scene is static does not affect the temporal flickering ranking among models. Ultimately, we selected a set of prompts, covering various topics, scenarios, and prompt lengths. Each prompt is accompanied by a prefix instructing the model to generate a static scene.

Motion Smoothness. Since *Subject Consistency*'s prompt suite involves movements performed by different subjects, they serve as a good benchmark for *Motion Smoothness* as well. To minimize the number of videos needed to be sampled for each model in evaluation, we share the same prompt suite for both dimensions.

Dynamic Degree. Considering the issue of the model tending to generate static videos even when prompted with descriptions of motion, we use the same prompt suite as *Subject Consistency*'s, which includes a variety of motion descriptions.

Object Class. We use the COCO dataset [68] and drop the object `mouse`, due to the potential confusion as it can be interpreted as both device and animal. We then append articles to the rest of the 79 objects and create a list of prompts related to different object classes.

Multiple Objects. We categorize COCO objects into various groups so that it will be reasonable for them to appear together. These categories include animals, indoor items, dining objects, bathroom items, and outdoor items. We then generate a list of prompts by composing objects within each category.

Human Action. From the Kinetics-400 dataset [56], we carefully extract a subset of 100 actions by considering both diversity and minimal overlaps in their meanings. Our approach involves selecting only the actions that are unique. For instance, within the category of actions related to playing musical instruments, we only keep those actions that are considered dissimilar in terms of human posture and actions. The resulting selection contains a wide spectrum of actions. Subsequently, we integrate each action in the form of "a person is doing something", and craft a list of human-centric action prompts.

Color. We select representative classes from COCO objects and establish the color scope of our prompt suite. On the selection of objects, we select objects that are unique in shape and similar objects. For example, "skateboard" and "surfboard" are excluded due to their similar shapes and potential wrong detection results by detection models. A similar criterion is applied to the color selection, we aim to select colors to include a broad spectrum while avoiding closely related colors. For example, "gold" and "yellow" are considered similar, therefore we only include "yellow" in our color scope. Our prompts are generated by combining each object with a few of their typical colors, and we only keep objects with more than three typical colors.

Spatial Relationship. We organize COCO objects into different groups so that it is natural for them to be composed in the same scene with each other. Some examples of the categories include personal items, animals, and sports-related items. Additionally, we define relationship categories to be "left and right" and "top and bottom". We then select relationships that are reasonable for the objects within each category, resulting in a list of prompts designed to describe spatial relationships between objects.

Scene. We use the same prompt suite as *Background Consistency*, as both requires prompts describing different general scenes.

Appearance Style. We select a list of sentences covering a wide range of scenarios and themes and also define our list of appearance styles. The styles are carefully crafted to ensure diversity. For example, we include the representative "Van Gogh style" and traditional "Ukiyo style" for the clear contrast in their color schemes, brushwork techniques, and overall aesthetic expressions. Each scenario description is then composed with a list of appearance styles to form the prompts.

Temporal Style. We carefully curate a diverse list of representative temporal styles to represent a broad spectrum of camera movement and temporal effects commonly employed in video production. Our selected temporal styles include variations in motion speed, camera perspective, and dynamic effects, aiming to present a comprehensive range of cinematic techniques. Each sentence for a scenario is then composed with a list of temporal styles.

Overall Consistency. We create a range of prompts, covering different content categories and scenarios such as "natural scenery", "fantasy and sci-fi", "character and fictional beings" *etc*., these prompts are of varied length, and we include both general and specific descriptions in our prompts.

H.2. Prompt Suite per Category

In Section 3.2 of the main paper on *Prompt Suite Per Category*, we employ LLM [130] as the first step to categorize the collection of human-curated prompts into eight content categories. The input template for the language model is shown in Table A3. The accuracy of classification is around 95%, and we manually go through each classified prompt to filter out 100 prompts for each content category.

Animal. These prompts focus on various animals and their behaviors in different environments, such as "a frog eating an ant", "a harbour seal swimming near the shore", and "a squirrel eating nuts". This prompt suite captures diverse

The assistant gives helpful, detailed, and polite answers to the user's questions. Please act as a language expert, able to choose one or more suitable categories from [`Animal, Architecture, Food, Human, Lifestyle, Plant, Scenery, Vehicles`] for the given text. Given the input text, you should return the answer without explanation. For example, if the input is [`A man eats hamburgers.`], the output tag format should be [`Food, Human`].
The given text is `Input text`.

Table A3. **Category Classification.** We employ LLM to determine the content categories of collected text descriptions.

species from domestic pets to wild animals in various activities, such as feeding, playing, or simply existing in their natural or adapted environments.

Architecture. We keep prompts that include various types of architecture, including the different types of buildings and structures, such as "the view of the Sydney opera house from the other side of the harbor", "illuminated tower in Berlin", and "a tree house in the woods".

Food. These prompts are diverse and all revolve around food and beverages. They range from specific dishes and preparation methods to more conceptual food art and eating scenarios. Examples include "Freshly baked finger-licking cookies", "A person slicing a vegetable", and "Close-up video of Japanese food".

Human. These prompts describe a wide range of human activities, interactions, and scenes, each focusing on specific individuals or groups engaged in various actions. Here are some examples: "A family wearing paper bag masks", "Boy sitting on grass petting a dog", "Group of people protesting", and "Father and son holding hands". Each of these prompts paints a vivid picture of human life, capturing diverse moments from daily activities to special events, professional settings to personal interactions.

Lifestyle. These prompts describe various indoor scenes and activities, covering a wide range of settings and situations. For instance, "Interior design of the bar section" and "Dog on floor in room" are simple everyday indoor scenes. Each prompt captures a specific aspect of indoor life, ranging from personal moments and family interactions to professional and leisure activities, reflecting the diversity of experiences within indoor lifestyles.

Plant. These prompts mainly focus on plants and trees. Here are some examples: "Video of an indoor green plant", "A coconut tree by the house", and "Variety of trees and plants in a botanical garden".

Scenery. These prompts describe various natural and urban landscapes, each capturing a distinct aspect of the environment. Here are some examples: "View of the sea from an abandoned building", "Aerial footage of a city at night", and "Scenery of desert landscape". Each prompt can be of natural settings like beaches and mountains, the structured scenery of agricultural lands, or urban environments.

Vehicles. These prompts depict various forms of transportation and related scenes, including various vehicles like trains, cars, buses, motorcycles, and boats in diverse settings ranging from urban streets to natural landscapes. Here are some examples: "A modern railway station in Malaysia used for public transportation", "Train arriving at a station", "Elderly couple checking engine of automobile", and "Helicopter landing on the street".

I. Human Preference Annotation

I.1. Human Annotation Procedures

Labeling Instructions. To systematically communicate with human annotators about labeling rules, we prepare a labeling instruction document for each of the 16 dimensions. Each labeling instruction document consists of several important elements. First, we introduce the labeling user interface (shown in Figure 4 of the main paper), including the two videos in comparison, the location of prompts and questions, the control for video playback and stop, and the three choices to make (*i.e.*, "A is better", "B is better", or "Same quality"). Second, we explain the dimension of interest. Since we want to verify the human alignment of VBench in each fine-grained dimension, we conduct the labeling of different dimensions separately. In each document, we elaborate on the definition of the current dimension, including aspects to consider or discard. For instance, for the *Subject Consistency* dimension, annotators are asked to only focus on the look of the main subject, and not to consider the degree of temporal flickering, or the video alignment with the text prompt, and many other irrelevant dimensions. Each aspect to consider or discard is illustrated by both text descriptions and examples of synthesized videos. Third, we categorize various scenarios that annotators may encounter while annotating this dimension (*e.g.*, what is considered as "better", and what is considered as "same quality"). For each scenario, we provide explanatory examples.

Quality Assurance in Preference Annotations. To guarantee the accuracy of human preference annotations, we implement a systematic five-step approach: ***1)*** Labeling Instructions Preparation: For each evaluation dimension, we provide clear and well-organized labeling instructions with examples. ***2)*** Pre-Labeling Trial: Prior to the main annotation task, we conduct a pre-labeling trial, where annotators are assigned to annotate only 60 samples. We go through all 60 annotations and communicate with annotators about each wrong label, and clarify any misunderstanding or potential doubts in the labeling instructions. ***3)*** Labeling In-

structions Update: We update the labeling instructions according to feedback from the human annotators, and supplement the wrongly labeled samples into the labeling instructions. ***4)*** Post-Labeling Checks by Annotators: Upon labeling all samples for a particular dimension, the samples are grouped as 60 samples per package. In each package of 60 samples, human annotators go through 20% of randomly selected samples for quality checking. If for any package the error rate exceeds 10%, the entire package is sent back for re-labeling conducted by a different annotator. ***5)*** Post-Labeling Checks by Authors: Upon labeling and possible re-labelings, we conduct the same post-labeling checks procedure similar to step 4. For any labeling errors spotted, we communicate with the human annotator for correction, and ask them to go through the entire package again. If any package reports an error rate higher than 10%, the entire labeled samples (all packages) for this dimension are considered invalid. We communicate with human annotators regarding possible problems encountered during annotation, and go back to step 1 to conduct annotation for this dimension all over again.

I.2. VLM Tuning

VLM Tuning. Our human preference annotations are mainly used to validate the alignment of VBench evaluations with human perception in each dimension. We show that these annotations have the potential to be used for broader applications, such as fine-tuning VLM models to align more closely with human preferences. VideoChat [64] uses a pre-trained video encoder and a large language model to align and train video VLM with instruction data. We organize data from VBench for fine-tuning, enhancing the model's ability to assess video generation. By utilizing instruction data of the VBench Dataset coupled with human preference tags, VLM, through its cognitive chain, is with its Chain of Thought to choose suitable evaluation metrics and furnishing scores in accordance. Upon giving the prompt into the VLM and asking: "`Give the appropriate evaluation metrics`", we obtain the dimensions to be evaluated from Chain of Thought, as shown in Table A4. We feed the derived metrics along with the video into the VLM, and ask two additional questions: "`Please observe the video and describe it based on the provided metrics.`" and "`Predict the assessment score of the video within these metrics.`" We show examples of before and after VLM Tuning in Figure A25

Data Preparation. We organize human preference annotations and model scores across different dimensions into training datasets. For model scores, each video's ratings are mapped from 0-1 to a 0-10 scale and are coupled with questions used for human annotation to form our instruction data. Evaluations of different aspects of the same video are grouped to facilitate the VLM's ability to engage in multi-turn dialogues. As for human preference annotations, pairs of videos along with corresponding questions and preference options are organized into instruction data, thereby somewhat enhancing the model with the capability to make evaluative judgments.

Implementation Details Our approach uses VideoChat-embed [64] as the baseline model. The model undergoes fine-tuning on a set of 30,000 instruction pairs. We fine-tune the model for 3 epochs using a learning rate of 2e-5 and an overall batch size of 64. The training for our model takes about one hour when performed on 8 A100-80GB GPUs.

`<SYSTEM>:` You are an assistant with fair and objective standards for video assessment. Given the input statement for the video generation model: `{input text}`, please select appropriate evaluation dimensions from `['subject consistency', 'background consistency', 'temporal flickering', 'motion smoothness', 'dynamic degree', 'aesthetic quality', 'imaging quality', 'object class', 'multiple objects', 'human action', 'color', 'spatial relationship', 'scene', 'temporal style', 'appearance style', 'overall consistency'].`

`<USER>:` a dilapidated phone booth stood as a relic of a bygone era on the sidewalk, frozen in time.
`<ASSISTANT>:` Based on the input statement, the following evaluation dimensions may be relevant: `subject consistency, background consistency, temporal flickering, motion smoothness, dynamic degree, aesthetic quality, imaging quality, scene, temporal style, overall consistency`

Table A4. **VLM Chain of Thought.** VLM suggests appropriate evaluation dimensions based on the user-input video prompt.

J. More Implementation Details

J.1. Video Generation Models in Evaluation

To evaluate our benchmark on recent advances, we adopt four models for comparison, and more will be added as they become open-sourced. Details of models and sampling

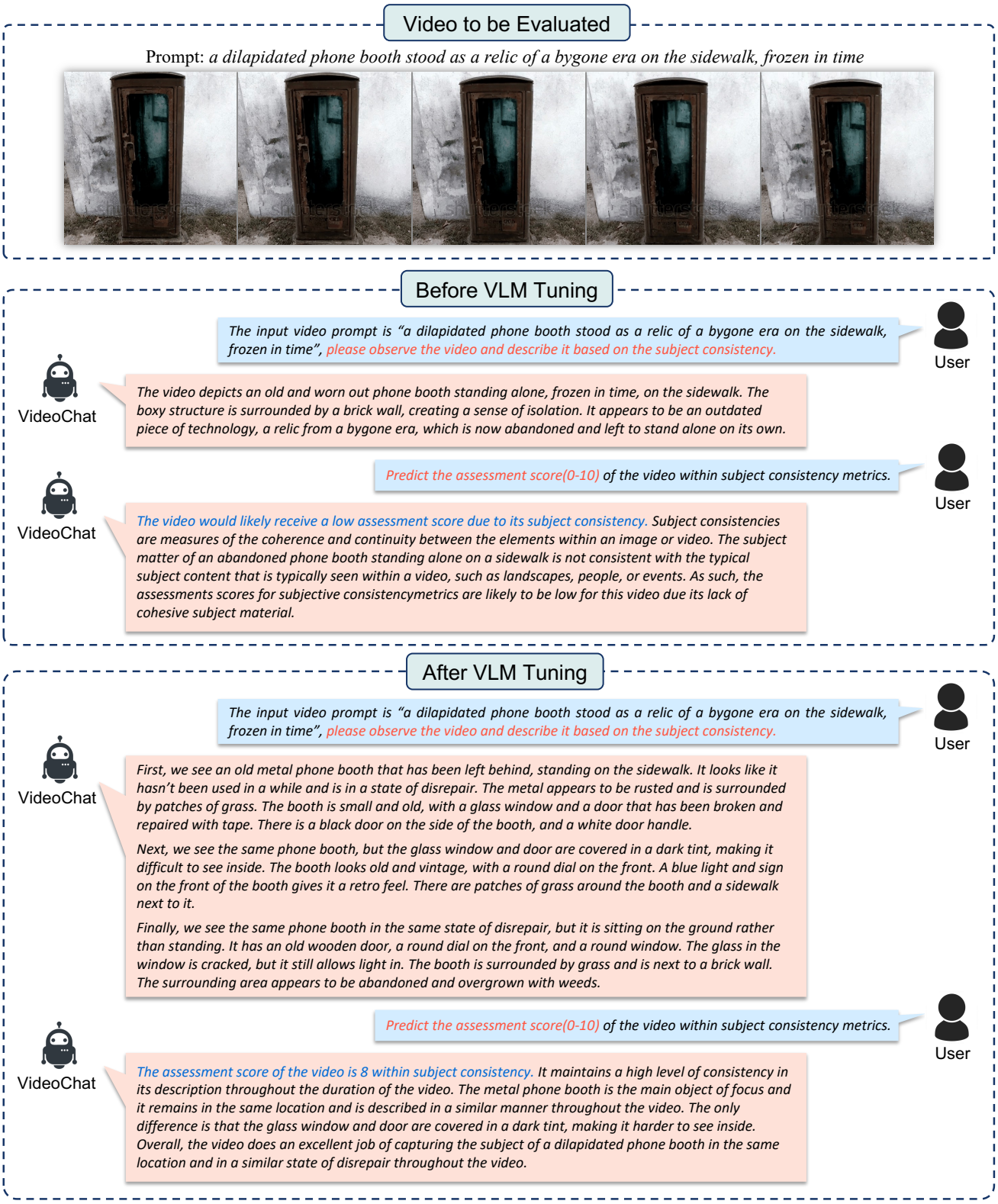


Figure A25. **VLM Tuning Example.** We use human preference annotation to fine-tune a VLM, and demonstrate that such fine-tuning can improve VLM's ability in evaluating video generation capabilities in specific dimensions.

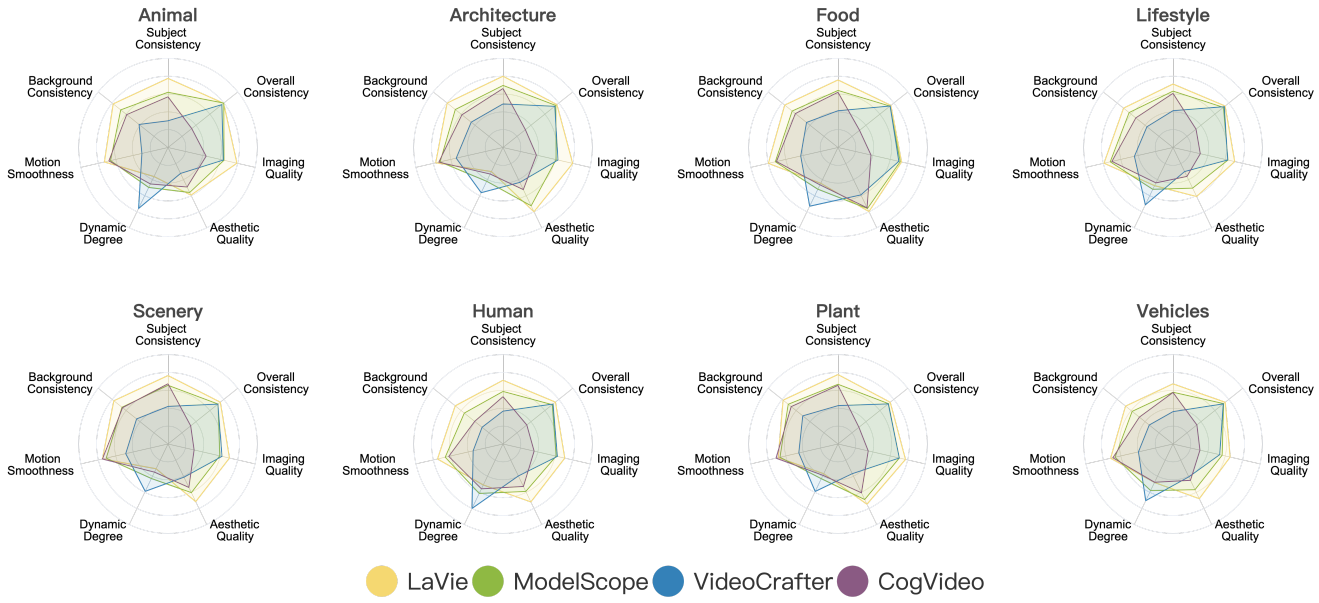


Figure A26. **VBench Results across Eight Content Categories (by Category per Chart)** (best viewed in color). For each chart, we plot the VBench evaluation results across different models on the same content category.

Table A5. **Validate VBench's Human Alignment.** We report *VBench Win Ratios (left) / Human Win Ratios (right)* for each dimension and each model. Our experiments show that VBench evaluations across all dimensions closely match human perceptions.

Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
LaVie [104]	**67.87% / 69.95%**	**85.27% / 65.04%**	**73.42% / 87.96%**	**69.54% / 65.65%**	41.81% / 53.10%	**77.56% / 83.41%**	**77.20% / 79.46%**	**57.55% / 79.20%**
ModelScope [72, 98]	49.07% / 56.30%	49.96% / 56.36%	65.42% / 62.44%	58.61% / 59.58%	52.92% / 53.84%	67.74% / 63.15%	60.00% / 68.53%	49.37% / 49.58%
VideoCrafter [35]	24.72% / 20.42%	15.89% / 27.21%	31.20% / 43.64%	10.00% / 13.80%	**68.47% / 62.18%**	35.34% / 32.33%	55.05% / 37.85%	54.18% / 41.77%
CogVideo [41]	58.33% / 53.33%	48.88% / 51.40%	29.96% / 5.96%	61.85% / 60.97%	36.81% / 30.88%	19.35% / 21.11%	7.74% / 14.16%	38.90% / 29.45%
Correlation	96.51%	94.80%	88.73%	99.80%	82.09%	98.65%	92.16%	80.37%

Models	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
LaVie [104]	53.37% / 57.97%	**54.43% / 58.13%**	**52.31% / 51.37%**	52.30% / 49.81%	**59.69% / 77.52%**	**61.85% / 58.22%**	**69.07%** / 55.73%	**70.82% / 77.35%**
ModelScope [72, 98]	**57.15% / 62.15%**	51.10% / 53.07%	50.12% / 49.73%	53.25% / 53.15%	48.22% / 50.00%	57.48% / 54.93%	65.40% / **57.50%**	66.31% / 60.07%
VideoCrafter [35]	48.74% / 49.63%	52.17% / 47.87%	48.71% / 47.92%	**56.11% / 54.66%**	52.79% / 46.05%	36.67% / 40.07%	65.40% / 51.90%	62.65% / 48.10%
CogVideo [41]	40.73% / 30.24%	42.30% / 40.93%	48.86% / 50.98%	38.33% / 42.38%	39.30% / 26.43%	44.00% / 46.78%	0.13% / 34.87%	0.22% / 14.48%
Correlation	98.98%	89.15%	60.73%	97.59%	94.07%	99.65%	97.53%	93.27%

Table A6. **VBench Evaluation Results on the WebVid-Avg Reference Baseline.** This table shows the VBench evaluation results on the *WebVid-Avg* baseline. We provide results from other models and baselines as well for a comprehensive view.

Models	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Appearance Style	Temporal Style	Overall Consistency
LaVie [104]	91.41%	**97.47%**	96.38%	49.72%	**54.94%**	**61.90%**	**23.56%**	**25.93%**	**26.41%**
ModelScope [72, 98]	89.87%	95.29%	95.79%	66.39%	52.06%	58.57%	23.39%	25.37%	25.67%
VideoCrafter [35]	86.24%	92.88%	91.79%	**89.72%**	44.41%	57.22%	21.57%	25.42%	25.21%
CogVideo [41]	**92.19%**	95.42%	**96.47%**	42.22%	38.18%	41.03%	22.01%	7.80%	7.70%
Empirical Min	14.62%	26.15%	70.60%	0.00%	0.00%	0.00%	0.09%	0.00%	0.00%
WebVid Avg	96.17%	96.59%	98.17%	44.13%	42.37%	58.22%	22.15%	25.77%	34.14%
Empirical Max	100.00%	100.00%	99.75%	100.00%	100.00%	100.00%	28.55%	36.40%	36.40%

Table A7. **VBench Results across Eight Content Categories.** We show the VBench evaluation results on the four T2V models, across eight content categories, on various evaluation dimensions.

Models	Categories	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Overall Consistency
LaVie [104]	**Animal**	97.49%	97.18%	97.29%	15.20%	48.26%	68.81%	**26.43%**
	Architecture	**98.04%**	**97.38%**	97.83%	5.20%	**54.20%**	**69.30%**	25.46%
	Food	97.11%	96.90%	**98.18%**	28.80%	54.15%	65.24%	24.88%
	Lifestyle	96.10%	96.19%	98.08%	33.60%	48.76%	64.02%	24.43%
	Scenery	97.27%	97.06%	97.58%	6.40%	51.76%	63.86%	24.56%
	Human	96.11%	95.88%	97.57%	**39.00%**	51.87%	64.07%	24.63%
	Plant	97.52%	97.20%	96.73%	16.40%	52.68%	67.86%	24.50%
	Vehicles	95.23%	95.82%	97.11%	34.00%	50.70%	61.02%	24.51%
ModelScope [72, 98]	**Animal**	94.08%	95.80%	96.40%	37.20%	47.32%	60.30%	**26.58%**
	Architecture	**95.77%**	95.88%	**97.20%**	24.80%	52.10%	58.38%	24.89%
	Food	94.53%	95.53%	97.17%	40.80%	**53.06%**	**64.39%**	24.40%
	Lifestyle	94.36%	95.17%	97.18%	41.00%	45.77%	59.62%	23.51%
	Scenery	94.88%	95.57%	97.03%	26.00%	48.57%	57.49%	23.28%
	Human	93.37%	94.21%	96.45%	**56.00%**	48.14%	58.41%	22.84%
	Plant	95.14%	**96.26%**	96.48%	26.40%	51.03%	63.83%	23.55%
	Vehicles	93.17%	94.61%	96.47%	50.20%	47.53%	55.75%	23.60%
VideoCrafter [35]	**Animal**	87.01%	92.40%	91.80%	79.60%	40.51%	59.79%	**25.47%**
	Architecture	**91.18%**	92.93%	**94.83%**	47.80%	43.71%	59.63%	24.27%
	Food	89.50%	92.87%	93.44%	75.00%	**48.19%**	63.47%	24.47%
	Lifestyle	89.51%	91.87%	93.63%	72.20%	39.84%	59.44%	24.01%
	Scenery	89.67%	92.86%	94.17%	51.80%	43.06%	58.98%	23.20%
	Human	88.50%	90.92%	92.35%	**86.20%**	42.62%	59.23%	23.31%
	Plant	89.86%	**93.57%**	93.72%	52.00%	41.81%	**63.81%**	23.41%
	Vehicles	88.38%	91.44%	93.04%	70.60%	42.95%	54.14%	23.39%
CogVideo [41]	**Animal**	92.95%	94.69%	96.65%	30.20%	45.37%	**48.45%**	8.26%
	Architecture	95.00%	94.65%	97.39%	10.20%	46.29%	45.33%	7.48%
	Food	94.08%	94.94%	96.99%	32.00%	**52.79%**	45.05%	7.01%
	Lifestyle	93.80%	93.93%	96.93%	28.00%	41.57%	41.28%	7.85%
	Scenery	**95.27%**	95.46%	**97.58%**	13.20%	46.72%	40.49%	7.66%
	Human	92.08%	92.29%	95.93%	**46.80%**	46.38%	43.81%	8.29%
	Plant	94.86%	**95.71%**	97.05%	19.60%	48.63%	43.22%	6.65%
	Vehicles	93.11%	93.27%	96.80%	33.60%	44.18%	41.05%	**8.34%**

strategy are listed as follows.

LaVie. LaVie [104] is a high-quality video generation model that incorporates cascaded latent diffusion models. Specifically, a set of temporal modules is attached to the vanilla Stable Diffusion [84] model and the entire model is jointly trained on both images and videos to achieve video generation. For each prompt, we sample 16 continuous frames of size 512×512 at 8 frames per second (FPS). We use the DDPM sampling of 250 steps. The initial random seed is set to 2 and the classifier-free guidance is set to 7.

ModelScope. ModelScope [72, 98] is a diffusion-based text-to-video generation model. We adopt its official inference code and sample 16 frames of size 256×256 at 8 FPS.

VideoCrafter. VideoCrafter [35] is a toolkit for text-to-video generation and editing. We adopt the VideoCrafter 0.9 version (*a.k.a.*, LVDM) and utilize its base generic text-to-video generation model. We use the official inference code to sample 16 frames of size 256×256 at 8 FPS. The initial random seed is set to 2 during sampling.

CogVideo. CogVideo [41] is a transformer-based text-to-video generation model that inherits the pretrained text-to-image model CogView2 [20]. Since the official inference code requires simplified Chinese input, we translate all prompts into Chinese. We sample 33 frames of size 480×480 at 10 FPS for each video, according to its default settings. During sampling, all stages are involved in the pipelines, including sequential generation, frame interpolation, and recursive interpolation. The initial random seed is also set to 2 for a fair comparison.

J.2. Reference Baselines

In the main paper, we devise the *Empirical Max* and *Empirical Min* baselines to approximate the maximum / minimum scores that videos might be able to achieve. We also devise the *WebVid-Avg* baseline to reflect the average video quality of WebVid-10M dataset [5] as a reference. The numerical

results are displayed in Table 1 in the main paper, and Table A6 in this Supplementary File. We provide additional details on approximating these values as follows.

Empirical Max. *(1) WebVid-10M's Maximum.* For dimensions where the 100% score is unlikely to be achieved by any video, we retrieve WebVid-10M's real videos and report the highest-scoring video's result. Examples of such dimensions include *Motion Smoothness*, *Scene*, *Appearance Style*, *Temporal Style*, and *Overall Consistency*. *(2) Theoretical 100%.* For dimensions where there exist videos that can achieve 100%, we directly use 100% as the empirical maximum value. For temporal consistency dimensions *Subject Consistency*, *Background Consistency*, and *Temporal Flickering*, a completely static video corresponds to the 100% score. For *Dynamic Degree*, a set of highly dynamic videos can achieve the 100% ratio of dynamic degree. For the frame-wise quality dimensions *Aesthetic Quality* and *Imaging Quality*, a video consisting of 100%-scoring frames results in a final 100% score. For video-text semantics dimensions *Object Class*, *Multiple Objects*, *Human Actions*, *Color*, and *Spatial Relationship*, videos with the correct semantics specified in the text prompt can score 100%.

Empirical Min. *(1) Gaussian Noise Videos.* For video-text feature similarity dimensions *Appearance Style*, *Temporal Style*, and *Overall Consistency*, we use videos of i.i.d. Gaussian noise and the corresponding prompt suites to compute the corresponding score, and select the smallest value as the approximated empirical minimum (with some actually reaching 0%). For *Temporal Flickering* and *Motion Smoothness*, we directly compute the score of the Gaussian noise videos and take the minimum scoring video's result. For *Human Action*, our method suite gives 0% on the Gaussian noise videos. *(2) Composed Videos.* For temporal consistency dimensions *Subject Consistency* and *Background Consistency*, we randomly sample frames from different WebVid-10M [5] videos to form a video with dynamically shifting content. This procedure is repeated 1000 times, and the minimum score among all videos obtained serves as the empirical minimum reference. *(3) Theoretical 0%.* For dimensions where there exist videos that can achieve 0%, we directly use 0% as the empirical minimum value. For *Dynamic Degree*, a set of static videos can achieve the 0% ratio of dynamic degree. For the frame-wise dimensions *Aesthetic Quality* and *Imaging Quality*, a video consisting of 0%-scoring frames results in a final 0% score. For video-text semantics dimensions *Object Class*, *Multiple Objects*, *Color*, *Spatial Relationship*, and *Scene*, videos with the incorrect semantics specified in the text prompt can score 100%.

WebVid-Avg. For dimensions where WebVid-10M videos can be retrieved with high confidence according to their captions, such as *Subject Consistency*, *Background Consistency*, *Motion Smoothness*, *Dynamic Degree*, *Aesthetic Quality*, *Imaging Quality*, *Appearance Style*, *Temporal Style*, and *Overall Consistency*, we compute the average score for all retrieved videos in relation to the corresponding dimension. This average score serves as a reference value for the average of real videos. The results are visualized in the main paper Figure 6 (b), and detailed in Table A6 in this Supplementary File.

J.3. Normalization for Radar Chart Visualization

In the radar charts, we perform normalization to clearly visualize the relative performance. We detail the normalization methods as follows:

- *Main Paper Figure 2. VBench Evaluation Results of Video Generative Models* - For each dimension, we map the maximum score achieved by one of the T2V models to 0.8, and the minimum score to 0.3, and linearly map the remaining models' scores to the radar chart axes. The radar chart axes have a range from 0.0 to 1.0.
- *Main Paper Figure 6 (a). T2V vs. T2I* - For each dimension, we map the maximum score achieved by one of the models (including T2I and T2V models) to 0.8, and the minimum score to 0.3, and linearly map the remaining models' scores to the radar chart axes. The radar chart axes have a range from 0.0 to 1.0.
- *Main Paper Figure 6 (b). T2V vs. WebV-Avg & Max* - For each dimension, we map the maximum score achieved by one of the models (including the *Empirical Max* and *WebVid-Avg* baselines) to 0.8, and the minimum score to 0.3, and linearly map the remaining models' scores to the radar chart axes. The radar chart axes have a range from 0.0 to 1.0.
- *Main Paper Figure 7. VBench Results across Eight Content Categories (by Model)* - For each dimension, there are 32 numerical results corresponding to the four T2V models and eight content categories. We map the maximum score among the 32 results to 1.0, and the minimum score among the 32 results to 0.0, and linearly map the remaining 30 scores to respective radar charts' axes. The radar chart axes have a range from 0.0 to 1.0.
- *Supp File Figure A26. VBench Results across Eight Content Categories (by Category)* - Unlike Figure 7 in the main paper which put different categories of the same model in one radar chart, in Figure A26 we use an alternative visualization method, that is, collecting different models' results of the same category in one radar chart. For each dimension, there are 32 numerical results corresponding to the four T2V models and eight content categories. We map the maximum score among the 32 results to 0.8, and the minimum score among the 32 results to 0.3, and linearly map the remaining 30 scores to respective radar charts' axes. The radar chart axes have a range from 0.0 to 1.0.

Table A8. **VBench Evaluation Results of Video vs. Image Generation Models.** We compare the performance of four video generation models against three image generation models. For each evaluation dimension, a higher score represents relatively better performance. For *Overall Consistency* we replaced the ViCLIP approach by CLIP to enable evaluating image generation models.

Models	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Overall Consistency
LaVie [104]	54.94%	61.90%	91.82%	33.32%	**96.80%**	86.39%	34.09%	52.69%	23.56%	32.96%
ModelScope [72, 98]	52.06%	58.57%	82.25%	38.98%	92.40%	81.72%	33.68%	39.26%	23.39%	31.99%
VideoCrafter [35]	44.41%	57.22%	87.34%	25.93%	93.00%	78.84%	36.74%	43.36%	21.57%	30.78%
CogVideo [41]	38.18%	41.03%	73.40%	18.11%	78.20%	79.57%	18.24%	28.24%	22.01%	27.80%
SD1.4 [84]	65.85%	**69.86%**	91.14%	34.39%	91.80%	90.57%	61.89%	52.33%	25.35%	32.59%
SD2.1 [84]	66.50%	69.10%	**93.42%**	51.22%	89.00%	**91.15%**	73.11%	**58.14%**	**25.48%**	33.08%
SDXL [81]	**70.38%**	68.79%	91.39%	**69.51%**	91.20%	88.92%	**86.17%**	54.65%	25.23%	**33.77%**

K. Potential Negative Societal Impacts

Video generation models could be maliciously applied to generate fake content involving human figures. Moreover, generative models can potentially inherit biases from the training datasets [21]. Therefore, we recognize the importance of considering ethical and safety aspects when evaluating video generation models. We plan to include safety and equality dimensions in future iterations of VBench. We also urge users to apply video generation models with discretion.

L. Limitations and Future Work

Limited Amount of Open-Sourced T2V Models: Currently, the number of open-sourced T2V models are still limited. We will open-source our VBench and encourage more T2V models to participate in the evaluation, including but not limited to [1–4, 8, 124], so that we can provide more informed insights into the current state of T2V, and provide more annotated data on T2V generation results generated by different models.

Evaluation of Other Video Generation Tasks: Text-to-video (T2V) is a fundamental task in video generation, and there are other related video generation tasks such as video-driven (*i.e.*, video editing) [11, 12, 18, 29, 43, 58, 62, 67, 69, 76, 80, 82, 101, 116, 119, 121, 127–129], image-driven (*i.e.*, image-to-video) [14, 16, 23, 27, 32, 76, 77, 88, 90, 100, 102, 106, 107, 122, 123], personalized video generation [33, 36, 51, 129], and other types of multi-modal-controlled video synthesis [15, 17, 42, 51, 58, 73, 75, 100, 103, 116, 117, 125, 126]. We build our VBench towards T2V as the initial step, and plan to extend our benchmark suite to accommodate other modalities' controls by adding towards the "*Video-Condition Consistency*" dimensions. Our "*Video Quality*" dimensions are readily available for evaluating these video generation tasks.

M. Additional Experimental Results

In this section, we provide additional numerical results that correspond to the main paper visualizations. We list the resulting tables and figures as follows:

- In Table A8, we show the VBench evaluation results of four video generation models and three image generation models, further illustrating through numerical results the significant differences that exist in certain dimensions between video generation models and image generation models (corresponding to *main paper Figure 6 (a)*). For *Overall Consistency* we replaced the ViCLIP approach by CLIP to enable evaluating image generation models.
- In Table A5, we show the win ratio on evaluation results predicted by VBench and Human across four models and all dimensions, along with the correlation (ρ) between Human and VBench results (corresponding to *main paper Figure 5*).
- In Table A6, we show the results of WebVid-Avg and compare them with the results of four models and other reference baselines (corresponding to *main paper Figure 6 (b)*).
- In Table A7, we show all the evaluation results of VBench across four models and eight different categories, providing numerical support for the relevant observations in the insights. (corresponding to *main paper Figure 7*). Additionally, for the *Dynamic Degree* dimension, intrinsic attributes of different categories naturally result in noticeable differences in the dynamic degrees among various categories. For instance, the `Human` category consistently exhibits the highest *dynamic degree* across different models. Conversely, the `Architecture`, `Scenery`, and `Plant` categories consistently showcase the lowest *dynamic degree* across various models, and the ascending order from lowest to highest remains consistent as `Architecture`, `Scenery`, and `Plant`. Due to this characteristic, the dynamic degree shows significant variability across different categories. Therefore, we isolate it as a supplementary dimension for additional analysis on top of other

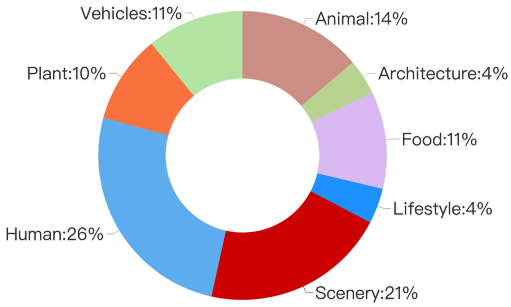


Figure A27. **WebVid-10M Dataset Categorical Distribution.** We visualize the percentage of data amount of each of the eight content categories in the WebVid-10M dataset.

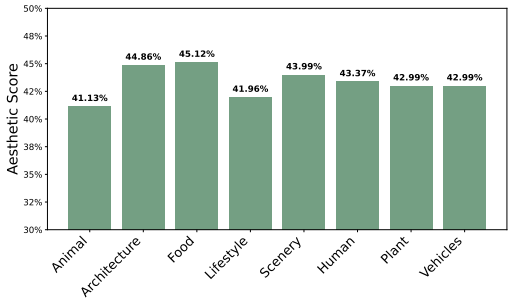


Figure A28. **Aesthetic Quality for Eight Categories in WebVid-10M dataset.** We visualize the aesthetic score of each of the eight content categories in the WebVid-10M dataset.

dimensions.

- In Figure A27, we show the statistical distribution of data amount of each of the eight content categories in the WebVid-10M dataset (supporting observations and insights mentioned in the *main paper Section 5*).
- In Figure A28, we show the aesthetic scores of eight different categories within the WebVid-10M dataset (supporting observations and insights mentioned in the *main paper Section 5*).