Video-Bench: Human-Aligned Video Generation Benchmark

Hui Han[1*] Siyuan Li[1*] Jiaqi Chen[2,3,4*] Yiwen Yuan[5*] Yuling Wu[5]
Yufan Deng[6] Chak Tou Leong[7] Hanwen Du[8] Junchen Fu[9] Youhua Li[10]
Jie Zhang[4] Chi Zhang[11] Li-jia Li[12] Yongxin Ni[13†]

[1]Shanghai Jiao Tong University [2]Stanford University [3]Fellou AI [4]Fudan University
[5]Carnegie Mellon University [6]Peking University [7]Hong Kong Polytechnic University
[8]Soochow University [9]University of Glasgow [10]City University of Hong Kong
[11]Westlake University [12]LiveX AI [13]National University of Singapore

https://github.com/Video-Bench/Video-Bench.git

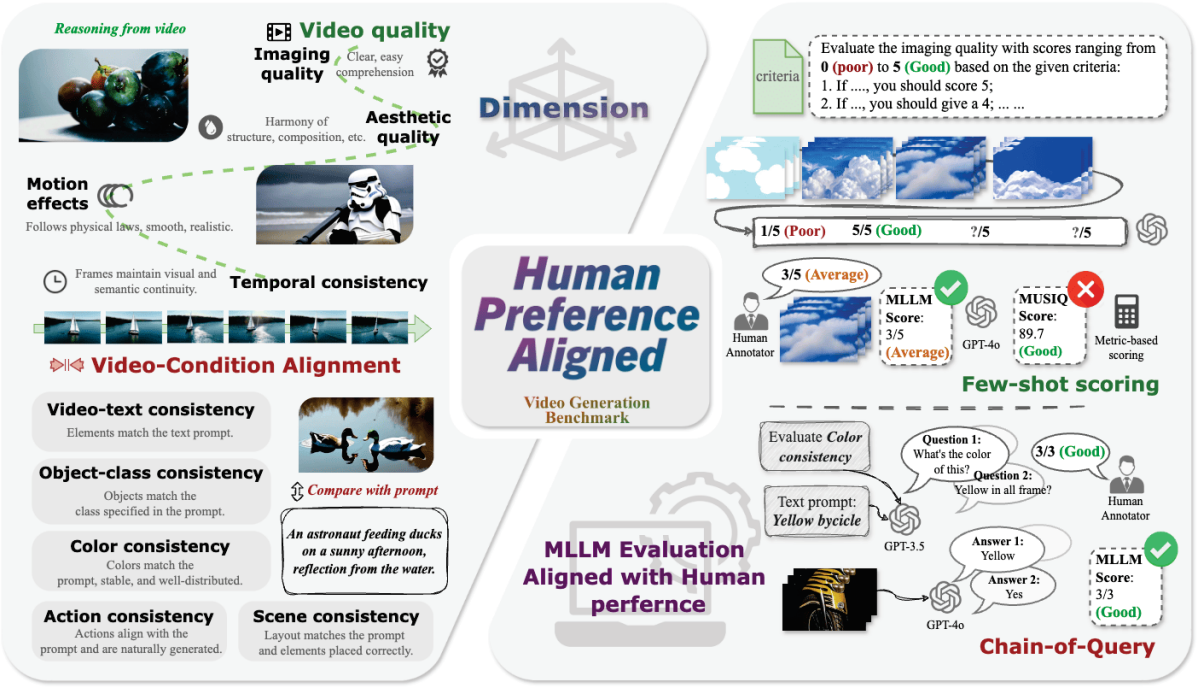

Figure 1. **Overview of `Video-Bench`.** *Left*: We introduce comprehensive evaluation dimensions in two main categories: ***video-condition alignment*** (Sec. 3.1.1) and ***video quality*** (Sec. 3.1.2). *Right*: For these two types of dimensions, we analyze their challenges (Sec. 4.1) and corresponding propose the MLLM-based automated video evaluation suite.

Abstract

Video generation assessment is essential for ensuring that generative models produce visually realistic, high-quality videos while aligning with human expectations. Current video generation benchmarks fall into two main categories: traditional benchmarks, which use metrics and embeddings to evaluate generated video quality across multiple dimensions but often lack alignment with human judgments; and large language model (LLM)-based benchmarks, though capable of human-like reasoning, are constrained by a limited understanding of video quality metrics and cross-modal consistency. To address these challenges and establish a benchmark that better aligns with human preferences, this paper introduces ***`Video-Bench`****, a comprehensive benchmark featuring a rich prompt suite and extensive evaluation dimensions. This benchmark represents the first*

*Equal Contribution.
†Corresponding author (email: niyongxin@u.nus.edu).

attempt to systematically leverage MLLMs across all dimensions relevant to video generation assessment in generative models. By incorporating few-shot scoring and chain-of-query techniques, Video-Bench provides a structured, scalable approach to generated video evaluation. Experiments on advanced models including Sora demonstrate that Video-bench achieve superior alignment with human preferences across all dimensions. Moreover, in instances where our framework's assessments diverge from human evaluations, it consistently offers more objective and accurate insights, suggesting an even greater potential advantage over traditional human judgment.

1. Introduction

In recent years, generative models have achieved remarkable advancements in the field of visual content generation [14, 21, 22, 40, 44, 46, 48]. These breakthroughs have further propelled progress in video generation [3, 15, 17, 38, 51, 57, 59, 63, 68], enabling unprecedented capabilities in creating dynamic and realistic videos from textual descriptions. As video generation models like Sora[1], Pika[2], and Runway[3], continue to evolve, there is an increasingly urgent need for reliable evaluation benchmarks to assess their quality. A reliable benchmark should be based on a comprehensive understanding of human preferences for video content: humans tend to prefer videos that are more aligned with input conditions such as color and scene descriptions (*i.e.*, high video-condition alignment), and those that demonstrate better aesthetic quality and temporal consistency (*i.e.*, high video quality).

To capture these aspects of human preferences, existing automated video generation benchmarks can be broadly categorized into two types: (1) ***Metrics and Embedding-based Benchmarks***: These benchmarks [23, 35, 36] attempt to combine various metrics to improve evaluation effectiveness. Metrics like Inception Score (IS) [49], Fréchet Inception Distance (FID) [16], Fréchet Video Distance (FVD) [55, 56] are used to calculate video quality [23]. Additionally, embeddings from pre-trained models such as CLIP [45] or BLIP [31] are utilized to measure video-text alignment. (2) ***Large Language Model (LLM)-based Benchmark***: LLMs has demonstrated promising results in evaluating text-to-image generation due to their strong language understanding and reasoning abilities [10, 11, 18, 27, 37, 66], leading a few video-generation benchmarks to integrate LLMs into their evaluation framework [52, 64].

However, both categories face different challenges in aligning with human preferences comprehensively. Metrics and embedding-based benchmarks, while providing quantitative measurements, often yield evaluation results that significantly misalign with human preferences [12, 41]. In contrast, LLM-based benchmarks, with their strong reasoning abilities, show promising potential in better understanding and simulating human evaluation logic, thus more closely imitating human evaluation patterns. Nevertheless, current LLM-based approaches encounter two critical limitations: In ❶ ***video-condition alignment*** evaluation, challenges arise in performing cross-modal comparisons between textual prompts and video content, In ❷ ***video quality*** evaluation, difficulties emerge in the ambiguity in converting textual critiques into concrete evaluation scores.

Overcoming these limitations is crucial for developing reliable evaluation benchmarks that authentically reflect human video preferences.

To address these challenges, this paper presents **`Video-Bench`**, a video generation benchmark shown in Figure 1 with comprehensive dimension suite and automatic multimodal LLM (MLLM) evaluation policy highly aligned with human preference. Building on this foundation, we propose a systematic MLLM-based framework to enhance the model's capability to directly comprehend video content and evaluate its quality. To ensure both depth and clarity in assessment, the protocol incorporates two key techniques: ❶ ***chain-of-query***, which facilitates iterative questioning to conduct a progressively multi-perspective evaluation along the preference dimension; and ❷ ***few-shot scoring***, which uses multimodal few-shot examples to guide the direction and scale of the evaluation process. Our experimental results show that Video-Bench evaluates video generation models without human intervention. Our proposed Chain-of-query and Few-shot scoring allows us to outperform all past evaluation methods by showing the highest correlation coefficient with human preferences. The following contributions are made in this paper.

- We present a comprehensive benchmark with rich dimensions, a large amount of prompt data, and precise manual annotations.
- We reformulate the video generation evaluation problem from the perspective of a MLLM-based framework, offering an alternative to traditional human evaluation and computational metrics, achieving a high level of alignment with human preferences.
- Our experimental results indicate that the proposed framework outperforms existing state-of-the-art evaluation methods.

2. Related work

2.1. Visual Evaluation with MLLM

In Liu *et al.* [34]'s summary, most LLM-based evaluation focus on measuring the alignment with respect to human instructions. GPT-4v Eval [70] uses LLM to do single-score grading and pair-wise comparison

[1] https://openai.com/index/sora/
[2] https://pika.art/
[3] https://runwayml.com/ai-tools/gen-2-text-to-video/

across four tasks. It shows that with prompting LLM's judgement has high agreement with human evaluation. LLMScore [37] transfers images into image-level and object-level descriptions and feeds the descriptions, along with original prompt into LLM to reason about the consistency. VIEScore [27] feeds rating instructions, considering semantic consistency and percentual quality aspects, prompt and image into MLLM and outputs the scores under each of the aspect. Many LLM-based evaluation focuses on question generation and answering (QG/A) [10, 18, 66]. TIFA [18] first creates binary questions from text prompt using LLM and then measures alignment by calculating the average number of correct answers produced by VQA system. VQ2 [66] also leverage VQA system but instead checks for whether the textual answer is accurate.

2.2. Video Generation Benchmark

Traditionally, T2V benchmarks follow either metric-based and model-based benchmarks for evaluation [23, 26, 35, 36]. For example, VBench [23] and EvalCrafter [35] both propose to evaluate generated videos by scoring them in multiple dimensions under the following categories–quality of video, motion consistency, temporal consistency and text-video alignment. Then the human ablation study is performed on those benchmarks. EvalCrafter [35] correlates user score with computed metrics by fitting a linear regression model, while VBench [23] calculates the Pearson and Spearman correlatoin between the two. The recent advancement of Large Language Models (LLM) and Multimodal Large Language Models (MLLM) has bring a promising direction to this field. More recent benchmarks, such as CompBench [52] and T2VScore [64] take a hybrid approach, only using MLLM to measure alignment with instructions, still using metrics-based methods to evaluate other aspects regarding perception.

3. Benchmark

3.1. Evaluation Dimension Suite

As shown in Figure 1, we divide "*video generation quality*" into two distinct aspects: "*video quality*" and "*video-condition alignment*". The former focuses solely on the quality of the video itself, while the latter assesses whether the video is generated according to the requirements specified by human instructions.

3.1.1. Video-condition alignment

The video-condition alignment is crucial for evaluating content quality, particularly in meeting specific user needs. Different consistency dimensions are scored based on difficulty, using either a three-point scale (1-3) or a five-point scale (1-5)[4].

[4]For more complex dimensions, the five-point scale is applied to encompass a broader range of criteria.

① **Object Class Consistency** This metric evaluates whether the objects presented in the video match those described in the text prompt. The focus is on whether the objects are generated correctly, are clearly identifiable, and whether their appearance and structure align with objective reality and human perception. In addition, the movement of objects should be examined for abnormal deformations. Scoring is based on a three-point scale.

② **Action Consistency** This metric assesses whether the actions in the video accurately reflect the descriptions in the text prompt. The focus is on the accuracy of generated actions, their clarity, and whether the appearance and process of the actions conform to objective reality and human cognition. Scoring is based on a three-point scale.

③ **Color Consistency** This metric measures whether the colors of objects in the video match those described in the text prompt. The colors in the video should remain consistent without sudden changes or discrepancies. Scoring is based on a three-point scale.

④ **Scene Consistency** This metric evaluates the alignment of the generated scene with the text prompt, ensuring that all relevant elements are clearly visible and arranged logically, with appearance and structure consistent with reality and human perception. Scoring is based on a three-point scale.

⑤ **Video-text Consistency** This metric assesses the overall consistency between the video and the text prompt, ensuring that all core elements (*e.g.*, humans, animals, actions, objects, scenes, style, spatial relationships, quantity relationships, *etc.*) are accurately represented and that the video quality does not impair user comprehension. Scoring is based on a five-point scale.

3.1.2. Video Quality

Video quality is a critical dimension for evaluating the visual fidelity of generated content, encompassing multiple aspects, each focusing on different elements of video quality. All dimensions are on a five-point scale (1 - 5).

① ***Imaging Quality*** This metric focuses on the technical quality of individual frames, assessing visual distortions such as noise, blur, overexposure, or other artifacts that may negatively impact viewer perception. The goal is to ensure that frames are technically as flawless as possible, minimizing defects that could affect overall quality.

② ***Aesthetic Quality*** Beyond technical considerations, the aesthetic quality of generated frames is evaluated to determine if they meet aesthetic standards and align with human perceptual expectations. This involves assessing artistic appeal, composition, and overall visual coherence to ensure that the generated content is both visually pleasing and naturally harmonious.

③ ***Temporal Consistency*** Temporal consistency is a crucial aspect of video quality that directly impacts the smoothness and naturalness of the video: *1) Visual feature consistency*: This metric ensures that visual features such as color, brightness, and texture transition smoothly across consecutive frames, avoiding abrupt changes or inconsistencies, thus maintaining continuity and high perceived quality. *2) Semantic consistency*: This aspect focuses on the consistency of objects, subjects, and scenes across frames, ensuring that objects maintain stable positions, shapes, and appearances, thereby avoiding sudden transformations or unnatural transitions that could detract from the viewing experience.

④ ***Motion Quality*** Motion quality is fundamental in distinguishing video from static images and plays a vital role in evaluating the dynamics of generated content: *1) Motion rationality*: This metric assesses whether the movement of objects within the video adheres to physical laws and appears realistic. Unrealistic or erratic movements can undermine the sense of immersion and significantly reduce perceived quality. *2) Motion amplitude*: This metric assesses whether the extent of movement is appropriate and aligns with the intended actions described in the prompt. Movements should neither be exaggerated nor overly subtle, ensuring that the generated dynamic content appears purposeful and lifelike.

These evaluation criteria rely on text-based guidelines to assess the quality of a video. However, the ambiguity in these textual guidelines and the lack of clear measurement scales make precise evaluation challenging.

3.2. Prompt Suite

Following the VBench [23], we designed a prompt suite for each evaluation dimension, with approximately 70-90 video generation prompts per dimension, with a total of 419 prompts. For "Action Consistency", "Temporal Consistency", and "Motion Quality", we combined human action data from the Kinetics-400 dataset [24] with rigid body and animal motion data from Subject consistency subset of VBench [23], to achieve a comprehensive and targeted evaluation of dynamic-related dimensions. For all other dimensions, our prompt suite aligns with the corresponding VBench [23] prompts. To mitigate bias from sampling randomness in video generation models, each prompt was sampled three times in our experiments.

4. Evaluation Framework with MLLMs

4.1. Challenges

Straight-forward prompting a multimodal language model (MLLM) to rate a video can present significant challenges due to the distinct complexities of video-condition alignment (Sec. 3.1.1) and video quality evaluation (Sec. 3.1.2) assessment. Here, we outline the two primary issues:

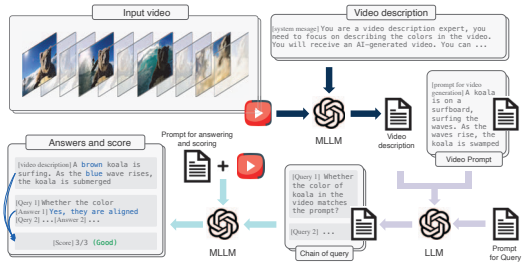

Figure 2. ***Chain-of-query*** **for video-condition alignment evaluation.** An iterative process where MLLM transforms video content into text descriptions, enabling detailed assessment of video-condition alignment through multi-turn queries.

- ❶ ***Video-condition alignment****: Cross-modal comparisons are difficult for MLLM.* When evaluating the consistency between generated videos and textual prompts, multimodal large language models (MLLMs) are required to compare visual signals with textual concepts. However, this cross-modal comparison is challenging [42], as MLLMs are prone to textual bias, often generating "hallucinations" and failing to accurately detect inconsistencies between text and video. MLLMs are faced with the task of comparing objects of different modalities. MLLMs are prone to hallucination, making incorrect statements about the user-provided image and the prompts [19, 67, 72].
- ❷ ***Video quality****: Textual critiques are ambiguous.* For dimensions such as video quality, the challenge lies in requiring an MLLM to assign an absolute score to a video based on human textual critiques. However, the criteria for such evaluations are inherently ambiguous. For example, both MLLMs and humans struggle to precisely distinguish between levels such as "very blurry", "somewhat blurry," and "very clear."

In our work, we propose ❶ ***chain-of-query*** and ❷ ***few-shot scoring*** to address above challenges.

4.2. Chain-of-query

For the ***video-condition alignment*** (Sec. 3.1.1) assessment, traditional LLM-based evaluation [52] relies on single-turn, direct question-answer methods, which often struggle with cross-modal alignment between video content and textual descriptions. This approach can miss critical details and fail to capture nuanced alignment, leading to incomplete evaluations.

To overcome this, we avoid direct cross-modal comparison by first transforming all relevant information from the video modality into a textual form. This allows for a more consistent, cross-modality comparison to assess alignment accurately. The evaluation

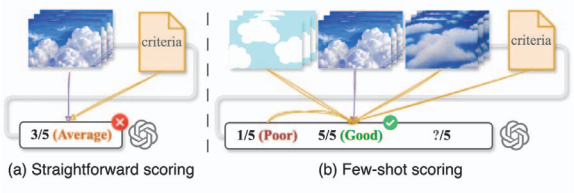

Figure 3. **Schematic scoring strategy comparison.** (a) Straightforward scoring assigns a single score based on the criteria, resulting in an average rating, while (b) Few-shot scoring leverages multiple examples for calibration, providing a nuanced assessment with scores ranging from poor to good.

process begins by generating an initial video description with MLLM based on the video prompt. Next, an LLM creates ***a chain of queries***, focusing on specific dimensions that need evaluation. These queries probe inconsistencies and overlooked aspects in the initial description. MLLM then revisits the video, addressing each query to refine the description with richer, dimension-relevant details. This iterative multi-turn process, as illustrated in Figure 2, results in a more comprehensive textual representation, enabling precise scoring for alignment. Specifically, our chain-of-query includes the following steps:

① **Video description:** MLLM is first prompted to generate a full description and a one-sentence summary of the video.

② **Query chain generation:** LLM generates N sets of questions with video description from first step and video generation prompt. The question generation strategy is predefined based on evaluation dimension. For example, for the color dimension. As shown in Figure 2, the first generated query is "*whether the color of koala in the video matches the prompt?*", while the second set focus on the color dominance with respect to other colors in the video: "*Has the koala's brown color ever been confused with the color of the waves?*".

③ **Answer chain generation:** MLLM is then prompted to answer the questions from above step. To ensure MLLM first analyze the whole video before answering individual questions, MLLM is first asked to perform reflection by re-generating a description. Finally MLLM responded "Yes, they are aligned" and "No, there is no confusion."

④ **Final scoring:** MLLM utilizes both the video content and the multi-turn conversation history, along with textual guidelines, to arrive at a final score.

4.3. Few-shot scoring

For ***video quality*** dimensions (Sec. 3.1.2) such as "Aesthetic Quality", our scoring criteria include descriptive text for each level, such as "3 points: moderate aesthetic quality" and "4 points: good aesthetic quality". Although each level has more detailed descriptions, the boundaries between them remain ambiguous. In our empirical observations, when we prompt an MLLM to evaluate specific dimensions like video quality, as shown in Figure 3 (a), the model often assigns the same (typically average) score to all videos. This outcome suggests that the MLLM is not sufficiently sensitive to variations in video quality, likely due to the inherent vagueness in textual criteria alone. We believe that textual guidelines alone are insufficient and that additional multimodal references are necessary to establish clearer benchmarks. To address this, we batch multiple videos generated from the same prompt, using each video as a contextual reference for the others within the same batch. Specifically, as illustrated in Figure 3 (b), when scoring the second video, all videos in the batch, as well as the score of the first video, serve as implicit references, providing a comparative framework that enhances the accuracy of quality assessment across the batch.

5. Experiment

5.1. Video Generation Models

We evaluate 4 open-source t2v models and 3 commercial models, spanning a range of advancements from earlier to more recent models. Specifically, we select 4 open-source models: LaVie [59], Show-1 [68], VideoCrafter2 [8], and CogVideoX [5] [65], as well as 4 commercial models: Pika-Beta [43]. Kling [28], Gen3 [47], and this diverse selection allows for a comprehensive evaluation of video generation techniques across different development stages, providing insights into the evolution of model performance.

5.2. Human preference annotation

We recruit 10 human annotators to manually rate each video[6]. For each generated video, we collect 4 human evaluation scores. The generated videos are shuffled and randomly assign to the pool of 10 evaluators. This leaves us a total of 35,196 evaluations. Both human and MLLM evaluations follow the same scale and guidelines. A human expert examines the score generated by human evaluators to gate-keep the quality of the evaluations. Table 3 shows the inter-rater alignment score produced by human rators, which is comparable to the human self-alignment scores produced by other studies [42].

5.3. Implementation details

Our experiment adopts GPT-4o (Version: `gpt-4o-2024-08-06`) and GPT-4o-mini (Version: `gpt-4o-mini`) as MLLM and LLM, respectively. GPT-4o handles multimodal inputs including texts and video frames, while GPT-4o-mini is only prompted with texts. The evaluation instructions

[5]CogVideoX-5B,the larger model in the CogVideoX series.

[6]The human annotators are provided with an extensive guide to ensure they provide quality evaluations of the videos given the prompts. The detailed guide is included in the Appendix.

Table 1. ***Video-Bench*** **Leaderboard.** Higher scores indicate better performance. The best score in each dimension is highlighted in bold. "Avg Rank" is the average rank of multiple dimensions, the lower the better.

Model	*Video quality*					*Video-Condition Alignment*						Overall
	Imaging Quality	Aesthetic Quality	Temporal Consist.	Motion Effects	Avg Rank	Video-text Consist.	Object-class Consist.	Color Consist.	Action Consist.	Scene Consist.	Avg Rank	Avg Rank
Gen3 [47]	**4.66**	**4.44**	**4.74**	**3.99**	**1**	4.38	2.81	2.87	2.59	**2.93**	2	**1**
Cogvideox [65]	3.87	3.84	4.14	3.55	3	**4.62**	2.81	**2.92**	**2.81**	**2.93**	**1**	2
VideoCrafter2 [8]	4.08	3.85	3.69	2.81	4	4.18	**2.85**	2.90	2.53	2.78	3	3
Kling [28]	4.26	3.82	4.38	3.11	2	4.07	2.70	2.81	2.50	2.82	5	4
Show-1 [68]	3.30	3.28	3.90	2.90	5	4.21	2.82	2.79	2.53	2.72	4	5
LaVie [59]	3.00	2.94	3.00	2.43	7	3.71	2.82	2.81	2.45	2.63	6	6
PiKa-Beta [43]	3.78	3.76	3.40	2.59	6	3.78	2.51	2.52	2.25	2.60	7	7

Table 2. **The human alignment score.** This score is measured by Spearman's rank correlation coefficient. Higher score indicates better performance. The best score in each dimension is highlighted in bold. In practice, ComBench* [52] is a reproduction version on our benchmark metrics.

Metrics	*Video quality*				*Video-Condition Alignment*				
	Imaging Quality	Aesthetic Quality	Temporal Consist.	Motion Effects	Video-text Consist.	Object-class Consist.	Color Consist.	Action Consist.	Scene Consist.
MUSIQ [25]	0.363	-	-	-	-	-	-	-	-
LAION [29]	-	0.446	-	-	-	-	-	-	-
CLIP [45]	-	-	0.260	-	-	-	-	-	-
RAFT [54]	-	-	-	0.329	-	-	-	-	-
Amt [33]	-	-	-	0.329	-	-	-	-	-
ViCLIP [60]	-	-	-	-	0.445	-	-	-	-
UMT [32]	-	-	-	-	-	-	-	0.411	-
GRiT [62]	-	-	-	-	-	0.469	0.545	-	-
Tag2Text [20]	-	-	-	-	-	-	-	-	0.422
CompBench [52]*	-	-	-	-	0.633	0.611	0.696	0.633	0.631
Ours	**0.733**	**0.702**	**0.402**	**0.514**	**0.732**	**0.735**	**0.750**	**0.718**	**0.733**

are written by humans, including task description, important notes, strategy, and output format. The full prompt of the evaluation process is included in the Appendix.

6. Main results

6.1. Comparison with existing evaluation methods

As shown in Table 2, we compare evaluation methods from previous benchmarks (*e.g.*, EvalCrafter [35], VBench [23], and ComBench [52]) on our prompt suite, calculating Spearman correlations with human ratings. Our MLLM-based evaluation achieves the highest correlation. Notably, ComBench [52] uses single-round video description and evaluation, while our Chain of Query mechanism enables multi-round interactions, improving Video-Condition Alignment by 0.093. This highlights our framework's ability to capture richer visual information and enhance cross-modal semantic comparisons. Overall, Video-Bench has stronger correlations to metric-based evaluation over other LLM-based methods across all dimensions.

6.2. Human preference alignment

As reported by Table 3, its agreement with human scores is on par with the inter-rater agreement among humans, with an average score of 0.52. To compare the distribution of the two evaluations on a larger sample, we calculate the mean difference after bootstrapping 1000 iterations over $100k$ score pairs sampled with replacement. The average absolute mean difference across dimensions between our proposed approach and human evaluations is 0.18. ***This shows Video-Bench's potential to replicate human judgments.*** We include the results of mean difference and its 99% confidence interval across dimensions in the Appendix.

6.3. MLLM evaluation *vs.* human evaluation

Table 3 shows that human evaluations have low agreement on semantic consistency-related dimensions, i.e.

Table 3. **Inter rater agreement degree (Krippendorff's α).** Higher score indicates better performance. "HU" stands for human, "HA" stands for ***Video-Bench*** and "GPT" stands for evaluations from single-GPT without chain-of-query, few-shot prompting or grid-view components.

Entities	*Video quality*				*Video-Condition Alignment*					**Avg.**
	Imaging Quality	Aesthetic Quality	Temporal Consistency	Motion Effects	Video-text Consistency	Object-class Consistency	Color Consistency	Action Consistency	Scene Consistency	
HU - HU	0.63	0.55	0.57	0.57	0.47	0.51	0.55	0.37	0.42	0.52
HU - GPT	0.51	0.42	0.45	0.35	0.47	0.50	0.49	0.37	0.11	0.41
HU - HA	0.61	0.54	0.48	0.48	0.50	0.52	0.54	0.40	0.43	0.50

Table 4. **Ablation study of component design.** Metric: Spearman's rank correlation coefficient. "Consist." denotes "Consistency".

Few Shot Scoring	*Video quality*				**Avg.**
	Imaging Quality	Aesthetic Quality	Temporal Consist.	Motion Effects	
	0.639	0.627	0.526	0.452	0.561
✓	**0.733**	**0.702**	**0.531**	**0.514**	**0.620**

(a) **Ablation study on few shot scoring.**

Chain of Query	*Video-Condition Alignment*					
	Video-text Consist.	Object-class Consist.	Color Consist.	Action Consist.	Scene Consist.	**Avg.**
	0.671	0.690	0.699	0.662	0.675	0.679
✓	**0.732**	**0.735**	**0.750**	**0.718**	**0.733**	**0.7336**

(b) **Ablation Study on chain of query.**

the five dimensions under Video-Condition Alignment. Nevertheless, this corresponds to the findings from EvalCrafter [35] that users tend to prioritize visual appeal over text-to-video alignment.

Due to individual differences among human evaluators (some may give high scores to videos with good visual effects, even if they don't meet the requirements), human consistency tends to be lower in semantics-related evaluations [39], while the results from the proposed strategy do not exhibit such trend. In fact, when incorporating evaluations from Video-Bench, the agreement improves in the Video-Text Consistency dimension. This demonstrates that ***our MLLM-based evaluation mitigates the perception bias commonly found in human assessments***.

6.4. Does MLLM generate stable evaluations?

Research [2] has shown that no LLMs deliver repeatable outcomes across different tasks in different benchmarks. If there is significant variance across identical runs, it reduces the reliability and validity of the benchmark. To this end, we analyze the stability of our benchmark to make sure the evaluations are stable and the benchmark is reliable. We run the experiments on Imaging Quality dimension three times with identical and deterministic configuration.

The results achieve a TARa@3 (Total Agreement Rate-answer across 3 runs) score of 0.67, meaning that 67% of the videos obtain the same rating from three repeated runs. In addition, to measure the scale of rating difference across the three runs, we calculate the Krippendorff's α as well, which achieves 0.867. This means that even though only 67% of the videos achieve the exact same rating, the agreement of the ratings across identical runs are is highly substantial. ***The results show that Video-Bench produces evaluations with high agreement across different runs.***

6.5. Robustness against small variations

A common issue found in traditional metric-based visual evaluations is that adding small perturbations hardly invisible to the human eye can easily fool the metrics [1]. To measure the robustness of our method against such small variations, we apply Gaussian blur to a subset of our videos and re-run our evaluation method. Under the video-text consistency dimension, we observe less than 5% relative percentage error. This showcases ***small variations in the input video that do not affect human judgment will not undermine MLLM evaluation either***.

6.6. Comparing *v.s.* rating paradigm

In video generation evaluations, both comparison-based [23] and rating-based [30] evaluations have been adopted by researchers. A popular type of comparison-based evaluation is arenas [13], where human evaluators pick the winner out of two choices. However, studies have found that both LLM [71] and humans [4] suffer from position bias in comparison-based evaluations, where both tend to favor the first item they see. Another major drawback of pairwise comparison is its high cost. To obtain the relative ranking across N models, N evaluations is needed for rating-based approach while $N \times (N-1)/2$ evaluations are needed for comparison-based approach. The quadratic complexity bottlenecks the benchmark's capability in evaluating a large number of models. On the other hand, rating-based approaches may fail to capture the subtle differences between videos, especially in the quality related metrics. To mitigate this

Table 5. **Evaluation results on different base models.** For each dimension, we randomly select 30 prompts for comparison.

Base models	*Video quality*				*Video-Condition Alignment*					**Avg.**
	Imaging Quality	Aesthetic Quality	Temporal Consist.	Motion Effects	Video-text Consist.	Object-class Consist.	Color Consist.	Action Consist.	Scene Consist.	
Gemini1.5pro	0.602	0.583	0.367	0.340	0.600	0.656	0.602	0.491	0.579	0.536
Qwen2vl-72b	0.586	0.51	0.535	0.353	0.576	0.669	0.634	0.637	0.600	0567
GPT-4o-0513	0.711	**0.690**	0.484	0.427	0.657	0.755	0.621	0.621	0.619	0.621
GPT-4o-0806	**0.807**	0.667	0.494	**0.469**	**0.750**	**0.767**	**0.676**	**0.761**	**0.73**	**0.680**
GPT-4o-1120	0.724	0.651	**0.538**	0.309	0.711	0.749	0.621	0.674	0.619	0.622

problem, we propose a few-shot scoring component where the evaluation model can have a sense of the relative quality across models. Table 4a shows that with few-shot scoring, the correlation with respect to human evaluations on average arises by 10.33% across different dimensions without increasing runtime complexity. This shows that the proposed rating paradigm can achieve ***high alignment with human evaluations*** while still being ***low-cost and free of position bias***.

6.7. Ablation study

Table 4 shows the ablation study on alignment with humans. We observe that our proposed components are all necessarily effective in reaching higher alignment with humans. Adding each component leads to a significant increase in the human-alignment score across all dimensions. Based on such observations, we validate that both components are effective in aligning with human preference. Experiments on Table 7 confirm that more reference videos lead to better performance.

6.8. Comparision on different base models

Table 5 indicates GPT-4o generally achieves superior video quality and alignment scores compared to Gemini1.5pro and Qwen2vl-72b, particularly in Imaging Quality (0.807) and Video-text Consistency (0.750) in GPT-4o-0806. However, performance does not consistently improve with newer GPT-4o versions. For instance, GPT-4o-1120 shows decreased Motion Effects (0.309 vs. 0.469 in GPT-4o-0806), suggesting potential regressions in temporal-motion detection across updates. Note that our benchmark results are recorded using the optimal version.

6.9. Simple *v.s.* complex prompt

As shown in Table 6, we tested state-of-the-arts (*i.e.*, Gen3, Kling and Pika) on both simple and complex prompts (*e.g.*, MovieGenBench), showing consistent performance across varying prompt lengths and complexities, demonstrating its robustness and versatility.

7. Conclusion

This paper has introduced Video-Bench, a human-aligned video generation benchmark leveraging evaluations from Multi-Modal Large Language Models (MLLMs). Extensive experiments and a human alignment study have demonstrated its advantages in efficiency , strong alignment with human preferences. In addition, we have provided insights into component design, emphasizing the potential for improving automatic evaluations through few-shot and chain-of-query technologies. Our work aims to support future research on video generation model development by providing a highly human-aligned benchmark using MLLMs in visual evaluation.

Table 6. **Performance on simple prompts *vs.* complex prompts.**

Prompt type	*Video quality*		*Video-Condition Alignment*	
	Imaging Quality	Motion Effects	Video-text Consistency	Action Consistency
Simple prompt	0.796	0.475	**0.749**	0.701
Complex prompt	**0.797**	**0.484**	0.725	**0.704**

Table 7. **Video quality comparison under different N-shot settings.**

Num.	*Video quality*				
	Imaging Quality	Aesthetic Quality	Temporal Consistency	Motion Effects	Avg Score
1	0.461	0.341	0.456	0.359	0.404
3	0.596	0.496	**0.529**	0.424	0.511
5	0.611	**0.511**	**0.529**	0.416	0.517
7	**0.624**	0.498	0.498	**0.557**	**0.545**

References

[1] Video Processing AI. Ways of cheating on popular objective metrics, 2024. Accessed: 2024-11-14.

[2] Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. Llm stability: A detailed analysis with some surprises. *arXiv preprint arXiv:2408.04667*, 2024.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[4] Niels J Blunch. Position bias in multiple-choice questions. *Journal of Marketing Research*, 21(2):216–220, 1984.

[5] Ljubiša Bojić. *Culture organism or techno-feudalism: how growing addictions and artificial intelligence shape contemporary society*. Institute for Philosophy and Social Theory, University of Belgrade, 2022.

[6] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[7] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024.

[9] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024.

[10] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

[11] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

[13] Hugging Face. Artificialanalysis/video-generation-arena-leaderboard. `https://huggingface.co/spaces/ArtificialAnalysis/Video-Generation-Arena-Leaderboard`. Accessed: 2024-11-14.

[14] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022.

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[16] 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. *Advances in neural information processing systems*, 30, 2017.

[17] 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.

[18] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[19] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[20] 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.

[21] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6080–6090, 2023.

[22] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023.

[23] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[24] 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.

[25] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[26] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024.

[27] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2024. *URL https://arxiv. org/abs/2312.14867*.

[28] Kuaishou. Kling ai model. Kuaishou Technology, 2024. Accessed: 2024-10-05.

[29] LAION-AI. Aesthetic predictor, 2025. Accessed: 2025-03-25.

[30] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[32] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023.

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[34] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. A survey of ai-generated video evaluation. *arXiv preprint arXiv:2410.19884*, 2024.

[35] 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.

[36] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] 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. *arXiv preprint arXiv:2303.08320*, 2023.

[39] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[40] 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.

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.

[42] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

[43] PikaLab. Pika. Company/Organization Name, 2024. Accessed: 2024-10-05.

[44] 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.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[47] Runway. Gen-3 model. Runway, 2024. Accessed: 2024-10-05.

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[49] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[50] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[51] 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.

[52] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

[53] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.

[54] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[55] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Syl-

vain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[56] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

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

[58] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024.

[59] 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.

[60] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[61] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3907–3916, 2024.

[62] 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.

[63] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[64] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.

[65] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[66] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

[67] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[68] 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.

[69] Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. Exploring perceptual limitation of multimodal large language models. *arXiv preprint arXiv:2402.07384*, 2024.

[70] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

[71] 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. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

[72] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.