VMBench: A Benchmark for Perception-Aligned Video Motion Generation

Supplementary Material

A. Human Perception Flow

Driven by the insights of the neuroscientific studies of motion perception, the human perception of motion within video can be systematically decomposed into two primary dimensions at a coarse level: the global parsing of motion fields and the capture of its finer details, as shown in Fig. 5. Specifically, the global perception of video motion fields, facilitates rapid evaluation of generated scenes plausibility by tracking macro-scale motion patterns, such as how smooth motion requires high frame rates to suppress temporal fragmentation (*e.g.*, jitter artifacts). Simultaneously, the fine-grained capture of motion in videos enables the detection of physically implausible movement patterns that violate fundamental physical laws, such as acceleration profiles violating Newton's laws or trajectories with positional discontinuities. The proposed VMBench systematically decomposes the two fundamental axes into granular perceptual criteria, thereby constructing a multi-dimension evaluation framework to quantitatively assess the spatiotemporal fidelity and motion coherence of generated videos.

B. Evaluation Dimension

B.1. Commonsense Adherence Score (CAS)

A prevalent issue in generated videos is the phenomenon that contradicts human perception and physical laws. As demonstrated in Fig. 6, generated videos frequently exhibit motions that defy physical laws and violate everyday intuitions and expectations, significantly compromising realism. Our CAS aims to evaluate whether generated videos align with human commonsense. As mentioned in the main text, we develop a specialized model to assess the commonsense quality of video content, categorizing it into five levels: Bad, Poor, Fair, Good, and Perfect.

First, we collect a comprehensive dataset of 10k generated videos from a wide range of sources. This dataset includes videos from legacy approaches as well as those generated by popular models [2–4, 15, 16, 38, 46, 91, 97]. The videos in our dataset come from two main sources: existing web datasets [48] and videos that we generate using these models. This approach ensures a diverse representation of video generation techniques and potential outcomes, capturing a wide spectrum of quality levels and possible commonsense violations. Such a comprehensive collection is crucial for training a robust evaluation model capable of assessing various aspects of video quality and realism. Second, we establish perceptual ground truth using VideoReward [48] to conduct systematic pairwise comparisons among the 10k videos. For each video pair, VideoReward determines which is preferable based on human perception standards. We then calculate a win rate for each video, representing its performance in all comparisons. These win rates are used to rank the videos, which are subsequently divided into five equal groups. Each group receives a label indicating its level of adherence to human commonsense expectations, from lowest to highest. Third, we choose the VideoMAEv2 [78] architecture for its temporal modeling capabilities, which are crucial for assessing commonsense adherence in video content. This model processes the input video and outputs logits for each of the five quality categories. We train VideoMAEv2 using the preference labels derived from the previous step. The model is initialized with a ViT-Giant [23] backbone pre-trained on large-scale video datasets. We fine-tune this model on our labeled dataset using 8 NVIDIA H20 GPUs. Our training process uses a batch size of 10, with input videos resized to $224 \times 224 pixels$. Each video clip consists of 16 frames, sampled at a rate of 4. We employ the AdamW optimizer with a learning rate of 1e-3 and weight decay of 0.1. The training schedule includes a 5-epoch warm-up period, followed by a total of 35 epochs. To enhance model performance, we implement layer-wise learning rate decay with a factor of 0.9 and a drop path rate of 0.3.

To compute the final CAS, we use a Mean Opinion Score (MOS) approach. The predicted probabilities for each class are weighted by their corresponding quality coefficients. The mapping function $G(i)$ converts the category index to quality weights as follows: $G(1) = 0(\text{Bad}), G(2) = 0.25(\text{Poor}), G(3) = 0.5(\text{Fair}), G(4) = 0.75(\text{Good})$, and $G(5) = 1(\text{Perfect})$. The CAS is then calculated using the formula provided in the main text:

$$\text{CAS} = \sum_{i=1}^{5} p_i G(i) \tag{1}$$

where p_i denotes the predicted probability for the i-th class. The resulting score provides a comprehensive measure of how well a generated video aligns with human expectations and commonsense understanding of the world.

B.2. Motion Smoothness Score (MSS)

Generated videos often exhibit blur and artifacts during object motion, particularly in areas with intricate details. This issue is especially pronounced when depicting complex movements that occur in the real world, as illustrated in Fig. 7. These visual inconsistencies likely stem from the

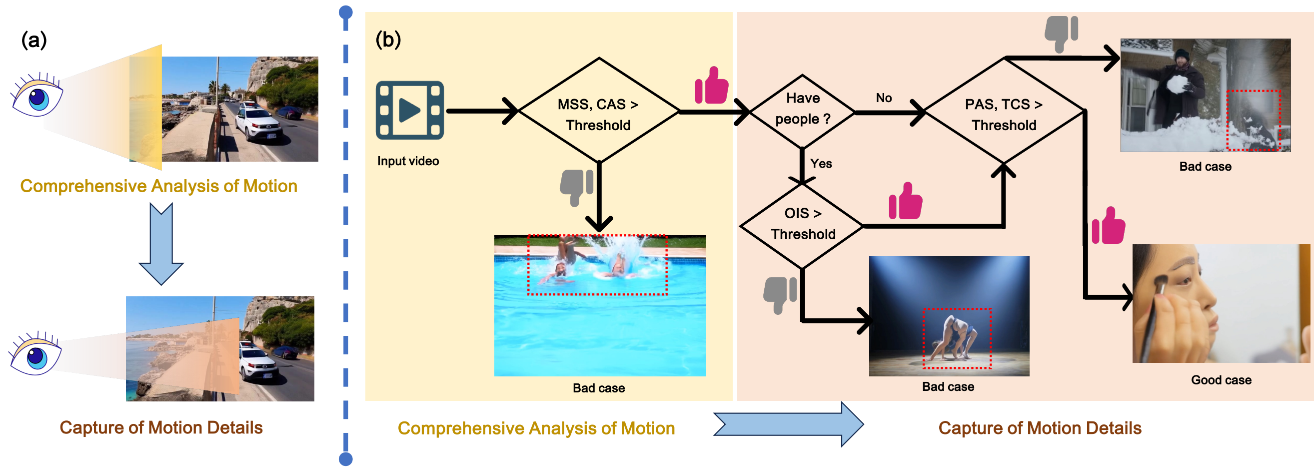


Figure 5. Our metrics framework for evaluating video motion, which is inspired by the mechanisms of human perception of motion in videos. (a) Human perception of motion in videos primarily encompasses two dimensions: Comprehensive Analysis of Motion and Capture of Motion Details. (b) Our proposed metrics framework for evaluating video motion. Specifically, the MSS and CAS correspond to the human process of Comprehensive Analysis of Motion, while the OIS, PAS, and TCS correspond to the capture of motion details.

model's difficulty in balancing the preservation of fine details with the representation of high-motion changes.

As mentioned in the main text, our MSS leverages Q-Align's [86] aesthetic score to detect artifacts. Here, we provide more details on how we quantify the frame-to-frame visual quality degradation magnitude ΔQ_t. The frame-to-frame visual quality degradation magnitude ΔQ_t is defined as:

$$\Delta Q_t = Q(f_{t-1}) - Q(f_t) \tag{2}$$

where $Q(f_t)$ represents the Q-Align aesthetic score for frame t. This formulation captures the change in visual quality between consecutive frames, with positive values indicating a decrease in quality. To determine the adaptive threshold $\tau_s(t)$, we conduct a statistical analysis of real video segments from datasets such as [14] and [55]. We analyze the relationship between motion amplitude and acceptable levels of quality degradation across diverse motion patterns. The threshold $\tau_s(t)$ allows for a higher tolerance of quality degradation in scenes with more intense motion. By incorporating this adaptive thresholding mechanism, our MSS effectively accounts for varying levels of acceptable blur in different motion scenarios, providing a more perceptually aligned evaluation of motion smoothness in generated videos.

The final MSS is computed as:

$$\text{MSS} = 1 - \frac{1}{T}\sum_{t=2}^{T}\mathbb{I}(\Delta Q_t > \tau_s(t)) \tag{3}$$

The MSS ranges from 0 to 1, where a score of 1 indicates perfect motion smoothness (no frames with significant quality drops), and lower scores indicate a higher proportion of frames with noticeable artifacts or blur.

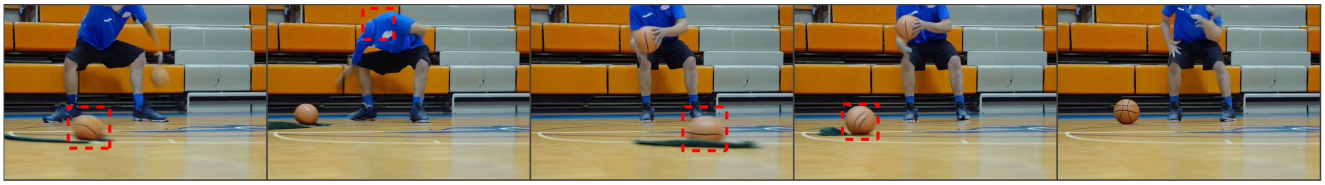
B.3. Object Integrity Score (OIS)

The integrity of moving objects in the generated videos is a crucial factor affecting the overall quality. Object integrity refers to the degree to which objects in the video maintain their physical structure and appearance consistent with real-world expectations. As illustrated in Figure 8, generated videos can sometimes exhibit abnormal distortions or deformations of moving objects. These distortions violate our perceptual expectations of normal object behavior and movement. We employ the MMPose toolkit [20] to detect key points of the primary subjects in the generated videos. These key points are then used to estimate the subjects' shapes in each frame. Our focus is on detecting perceptual issues (*e.g.*, distorted shapes) that are readily noticeable to the human visual system.

For a comprehensive anatomical analysis, we consider both length and angle variations of object components. Let $K = k_1, k_2, ..., k_n$ be the set of key points detected in each frame. Through statistical analysis of our datasets, we establish thresholds τ_L and τ_θ to detect changes in unnatural shape in lengths and angles, respectively.

For length analysis, we calculate the Euclidean distance $L_{i,j}(t)$ between connected key points k_i and k_j in each frame t. We then observe the variations in these lengths across frames, identifying potential distortions when changes exceed the threshold τ_L:

$$D_L(i,j) = \sum_{t=2}^{T}\mathbb{I}(|L_{i,j}(t) - L_{i,j}(t-1)| > \tau_L) \tag{4}$$

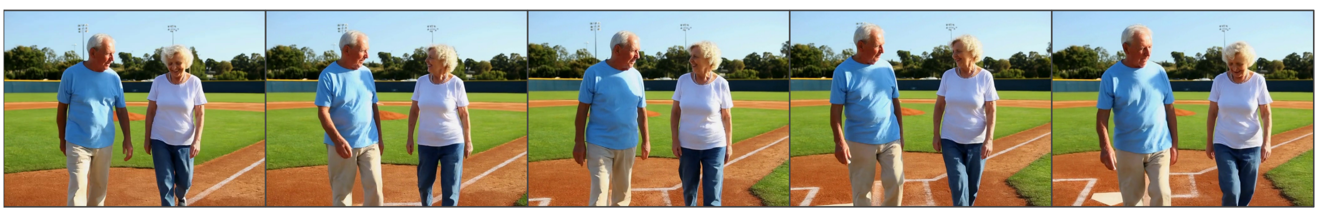


(a) violates the laws of physics (score 7%)

(b) naturalness movement (score 85%)

Figure 6. Visualization of Commonsense Adherence. (a) The ball exhibits perpetual rolling motion on the ground without external forces, violating physical laws and contradicting human perception. (b) All objects demonstrate motion consistent with natural physical principles.

(a) unsmooth motion (score 40%)



(b) smooth motion (score 90%)

Figure 7. Visualization of Motion Smoothness. (a) Both subjects exhibit significant blur during walking, with the female's facial features particularly affected, resulting in a loss of fine details. (b) Both subjects demonstrate fluid motion, with clear visibility of bodily details.

where $D_L(i,j)$ denotes the distortion count for the component between keypoints k_i and k_j, T represents the total number of frames, and $\mathbb{I}(\cdot)$ is the indicator function.

Similarly, for angle analysis, we compute the angles $\theta_{i,j,k}(t)$ formed by adjacent key points in each frame. We monitor these angles for abrupt changes that surpass the threshold τ_θ:

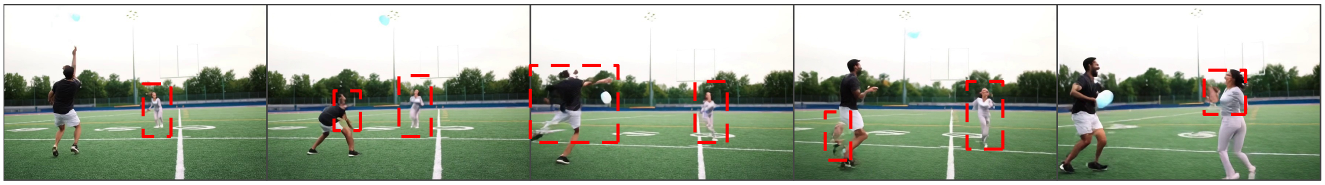
$$D_\theta(i,j,k) = \sum_{t=2}^{T} \mathbb{I}(|\theta_{i,j,k}(t) - \theta_{i,j,k}(t-1)| > \tau_\theta) \quad (5)$$

These length and angle analyses contribute to the compound anatomical deviation $\mathcal{D}_f^{(k)}$ for each anatomical component k in frame f. We establish tolerance thresholds $\tau^{(k)}$ for each anatomical component through statistical analysis of natural motion samples from datasets such as [14, 55, 88].

The OIS is then computed as:

$$\text{OIS} = \frac{1}{F \cdot K} \sum_{f=1}^{F} \sum_{k=1}^{K} \mathbb{I}\left(\mathcal{D}_f^{(k)} \leq \tau^{(k)}\right) \quad (6)$$

This formulation checks if the compound anatomical deviation $\mathcal{D}_f^{(k)}$ is within the acceptable threshold $\tau^{(k)}$ for each frame and anatomical component. The indicator function returns 1 for each instance where the deviation is within the threshold. We sum these values across all frames and anatomical components and then normalize by dividing by the total number of checks performed $(F \cdot K)$.



(a) distort shape during motion (score 20%)

(a) integrity object shape (score 80%)

Figure 8. Visualization of Object Integrity. (a) Both subjects exhibit varying degrees of bodily distortion, with their limbs becoming difficult to discern due to severe warping. (b) Both subjects maintain normal anatomical structure throughout the sequence, displaying no unnatural deformations.

(a) slightly camera motion (score 1%)

(b) medium camera motion (score 62%)

Figure 9. Visualization of Camera Motion. (a) The object and background remain relatively static, indicating subtle camera movement. (b) The scene exhibits noticeable changes, demonstrating a panning or tracking camera movement.

B.4. Perceptible Amplitude Score (PAS)

Motion amplitude in videos stems from two sources: camera motion, as illustrated in Fig. 9, and subject motion, as demonstrated in Fig. 10. Our PAS focuses on the latter. Traditional methods like RAFT [74] can be affected by camera motion when detecting subject movement. However, our approach effectively isolates subject motion from camera movement, enabling a more accurate perception of the primary subject's motion regardless of camera dynamics.

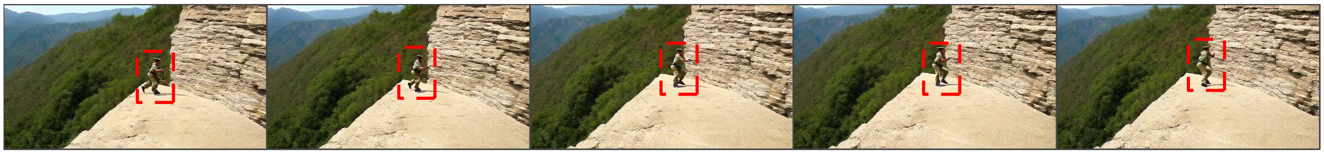
Our method begins by employing GroundingDINO [49] to detect the primary moving subject in the video, followed by GroundedSAM [63] to generate precise masks for this subject across frames. We then utilize CoTracker [36] to track key points for the main subject using these masks.

The motion magnitude is computed based on the average displacement of these key points. For each tracked key point p at frame t, we calculate its displacement as:

$$D(p^t) = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \tag{7}$$

The frame-level motion amplitude $\bar{D}_t$ is then calculated as the average displacement across all tracked key points for active subjects in frame t:

$$\bar{D}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} D(p_i^t) \tag{8}$$



(a) slightly subject motion (score 5%)

(b) high subject motion (score 85%)

Figure 10. Visualization of Subject Motion. (a) The main subject exhibits only minor changes throughout the video, indicating limited movement. (b) The subject completes a full range of actions, even moving out of frame, demonstrating a significant magnitude of movement.

(a) inconsistent subject appearance (score 0%)

(b) consistent subject appearance (score 100%)

Figure 11. Visualization of Temporal Coherence. (a) The female disappears and reappears throughout the video, while the male exhibits discontinuous behavior. (b) Both subjects maintain consistent presence and stability throughout the sequence, demonstrating superior temporal continuity.

where N_t is the number of tracked key points in frame t. To account for the context-dependent nature of human motion perception, we derive a set of perceptual motion magnitude thresholds τ_s for various scenarios s through statistical analysis of existing video datasets [14, 55]. These thresholds serve as the foundation for computing a motion score for each video. The Perceptible Amplitude Score (PAS) is then computed as:

$$\text{PAS} = \frac{1}{T}\sum_{t=1}^{T}\min\left(\frac{\bar{D}_t}{\tau_s}, 1\right) \tag{9}$$

where T is the total number of frames in the video, $\bar{D}_t$ is the frame-level motion amplitude, and τ_s is the perceptual motion threshold for scenario s. This method ensures that the PAS accounts for both the magnitude of motion and its perceptual significance in different contexts, providing a more nuanced evaluation of motion in videos.

B.5. Temporal Coherence Score (TCS)

In generated video sequences, moving subjects often exhibit phenomena of sudden disappearance or appearance, as illustrated in Fig. 11. These temporal discontinuities significantly impact the perceived quality of motion. Stable temporal coherence is crucial for achieving high-quality motion in generated videos.

We employ GroundedSAM2 [62] for pixel-accurate instance segmentation and tracking across frames, maintaining persistent object IDs throughout the whole sequence. For objects exhibiting discontinuous existence, we apply a secondary verification phase using CoTracker [36] to track dense key points on target objects and construct their motion trajectories.

We then analyze these motion trajectories to determine whether any anomalous phenomena are present. Our approach mitigates false cases caused by legitimate object discontinuity through a rule-based filtering mechanism. These rules account for common scenarios, including: 1) Objects reappearing after occlusion or disappearing behind obstacles. 2) Objects entering or exiting frame boundaries. 3) Apparent size changes due to depth perception, such as objects appearing larger when moving closer or smaller when moving farther away. Let N be the total number of object instances in the video. For each object instance i, we define: $\mathcal{A}$: An indicator function that equals 1 if the object exhibits discontinuous existence, and 0 otherwise. $\mathcal{R}$: A function that validates legitimate transitions based on our rule-based filtering mechanism. It returns 1 if the transition is legitimate (i.e., falls under one of the three scenarios mentioned above), and 0 otherwise. The TCS is then computed as:

$$\text{TCS} = 1 - \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\mathcal{A}_i \wedge \neg \mathcal{R}) \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition inside the parentheses is true, and 0 otherwise. The term $\mathcal{A}_i \wedge \neg \mathcal{R}$ identifies objects that exhibit discontinuous existence ($\mathcal{A}_i = 1$) and do not have a legitimate reason for this discontinuity $\mathcal{R} = 0$. TCS ranges from 0 to 1, where a score of 1 indicates perfect temporal coherence (no anomalous discontinuities), and lower scores indicate a higher proportion of unjustified object vanishing or emerging events. This formulation ensures that the TCS accounts for both the presence of discontinuities and the legitimacy of these discontinuities based on our rule-based filtering, providing a nuanced evaluation of temporal coherence in videos.

C. MMPG

C.1. Prompts Statistic

In this section, we conduct Motion Prompts Statistics (as shown in Fig. 12) to emphasize VMBench's focus on motion. In Table (a), we perform a statistical analysis to demonstrate the superiority of our prompts compared to previous works, focusing on the number of prompts (NP), the number of motion prompts (NMP), the average length of prompts (ALP), the types of motion subjects (TS), place (TP), and actions (TA). We find that VMBench provides the most comprehensive coverage of action types and the most detailed prompt descriptions, making it an effective benchmark for evaluating the dynamic motion generation capabilities of video generation models. Fig. 12 (b) illustrates the distribution pattern of our motion prompts. It is evident that our prompts, while covering six major motion patterns, are particularly rich in content related to the most common mechanical and biological motions found in everyday life. This aligns with the characteristic of our prompts being realistic and sensible descriptions. Fig. 12 (c), Fig. 12 (d), and Fig. 12(e) respectively demonstrate the richness of subjects, places and actions within the prompts, highlighting the richness and variety of motion content. Fig. 12 (f) presents a well-distributed range of prompt lengths, and Fig. 12 (g) shows the distribution of motion subjects, reflecting the diversity among subjects in our prompts. We employ the dynamic evaluation method from DEVIL [45] to assess the dynamic grade of our prompts, as shown in Fig. 12 (h). The results indicate that our prompts exhibit a high level of dynamism overall, which poses a challenge for large models.

C.2. Human-LLM Reasoning Validation

To ensure that the prompts generated by the GPT-4o describe motion that exists in real life, we combine the efforts of both LLMs and humans to evaluate the plausibility of the prompts. We first utilize the strong reasoning capability of DeepSeek R1 [21] to evaluate the realistic reasonableness of motion descriptions logically (see Fig. 13), which results in a quantified score. After filtering out prompts with lower plausibility scores, we then recruit evaluators to verify the real-world validity of the prompts through a survey (as shown in Fig. 14). After a rigorous review process, we ultimately retain 1050 prompts that describe reasonably realistic motion.

D. Implementation Details

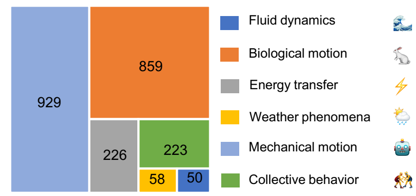
D.1. Inference Details of Video Generation Models

To ensure a fair comparison, we utilize the best open-source architectures and weights available for each model and maintain the optimal hyperparameters (including video resolution, sampling steps, scale, etc.) as demonstrated in their respective demos to generate the corresponding videos of approximately 5 seconds. Additionally, we record the time cost of model inference (excluding model loading) for reference. We list the inference details for each model as follows:

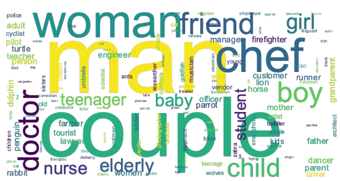
HunyuanVideo [38] The preset video resolution is 624×832 with a length of 129 frames. Using a 4-GPU parallel inference setup, the generation time for a single video is approximately 610 seconds.

Benchmark	NP	NMP	ALP	TS	TP	TA
VBench	800	246	6.69	125	58	132
EvalCrafter	700	228	12.93	169	63	237
T2V-CompBench	**1400**	962	10.64	308	167	545
FETV	619	422	11.27	168	101	334
DEVIL	810	481	16.03	270	296	556
VMBench	1050	**969**	**26.17**	**340**	**390**	**1216**

(a) Statistical analysis with other benchmark



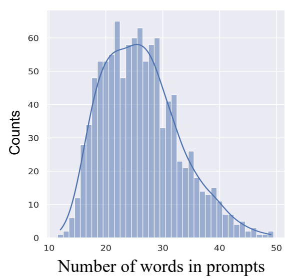
(b) Coverage of types of motion patterns.



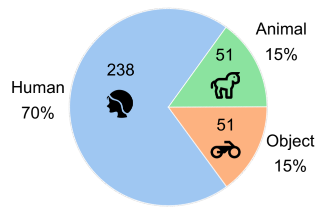
(c) Word cloud of subjects

(d) Word cloud of places

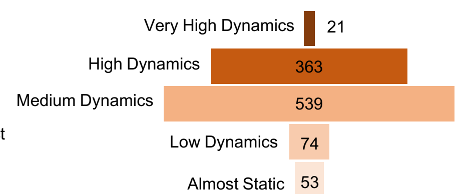
(e) Word cloud of actions



(f) Word counts of prompts



(g) Distribution of subjects



(h) Distribution of dynamic grade

Figure 12. Statistical analysis of motion prompts in VMBench. (a-h): Multi-perspective statistical analysis of prompts in VMBench. These analyses demonstrate VMBench's comprehensive evaluation scope, encompassing motion dynamics, information diversity, and real-world commonsense adherence.

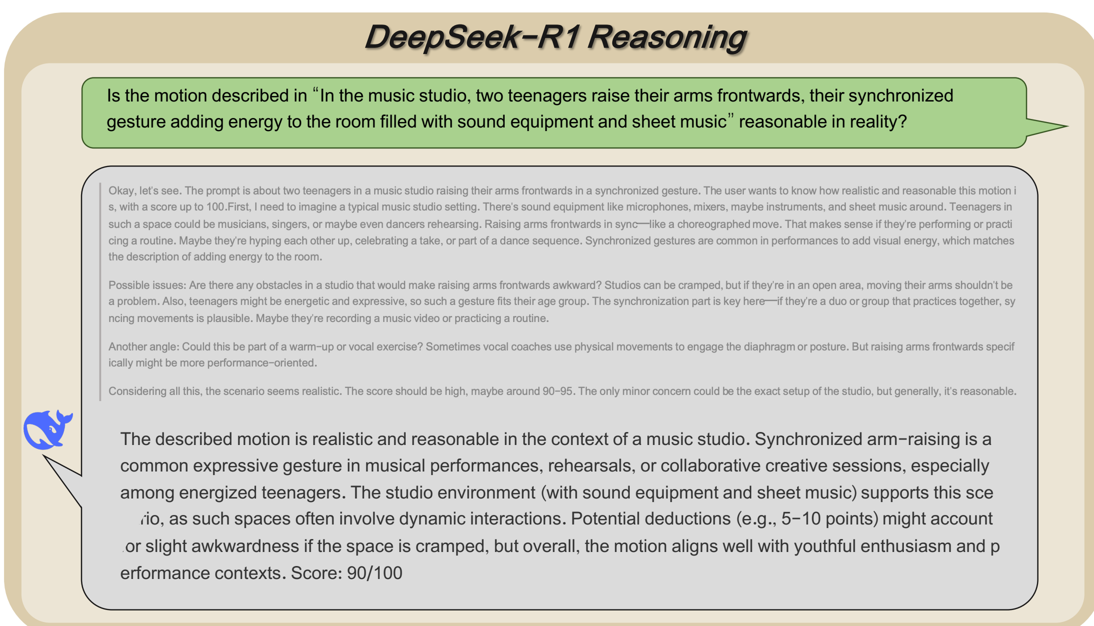


Figure 13. An Example of DeepSeek-R1 Reasoning. A case of evaluating the realistic reasonableness of a prompt using DeepSeek-R1.

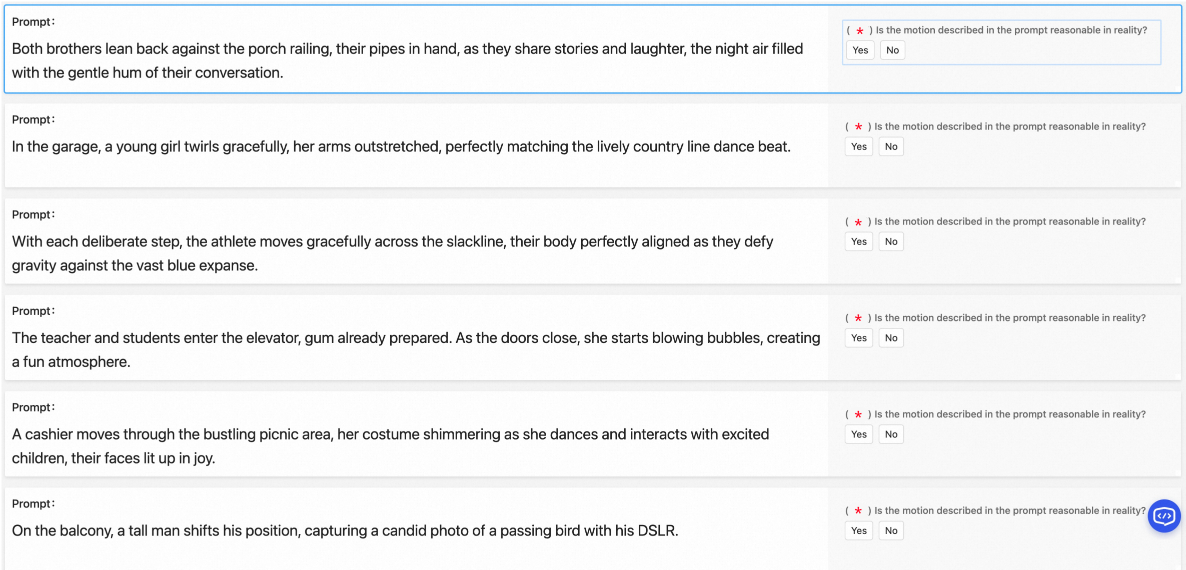


Figure 14. Manual Review of Prompt Validity in Real-World Scenarios. Some cases of manually reviewing the real-world validity of prompts.

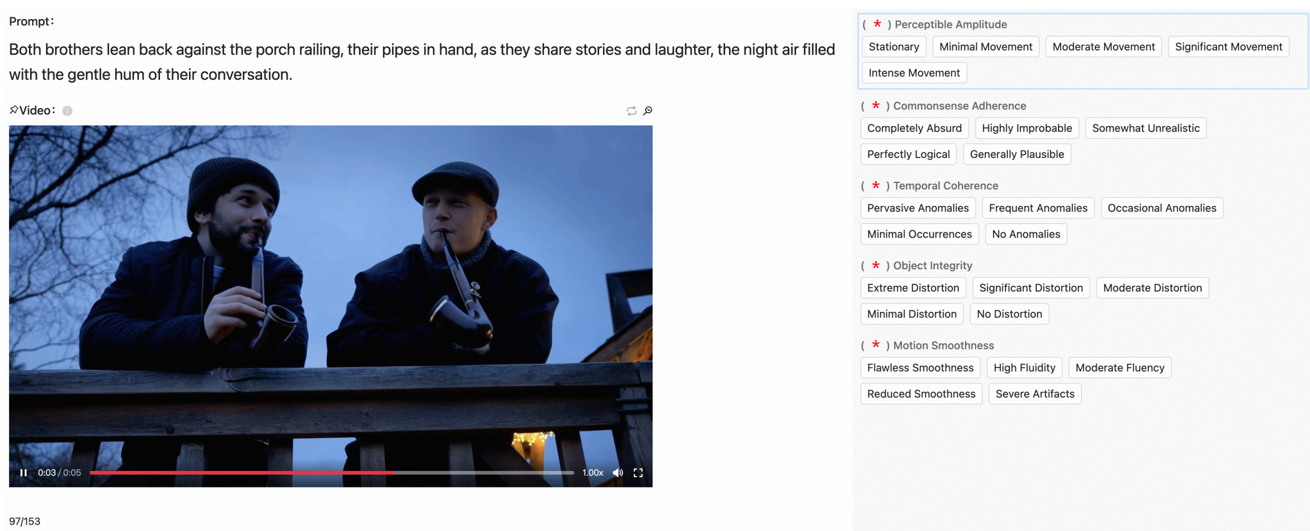


Figure 15. Human Annotation Procedure. Three annotators independently evaluate each aspect, re-watching the video for each question. Annotators are instructed to focus solely on the specific aspect being evaluated, disregarding other potential influences.

OpenSora [97] We use the Open-Sora v1.2 model version. The preset video resolution is 720×1280 with a length of 102 frames and uses 30 sampling steps. Using a 4-GPU parallel inference setup, the generation time for a single video is approximately 85 seconds.

CogVideoX [91] We use the CogVideoX-5B model version. The preset video resolution is 480×720 with a length of 49 frames. Using a 2-GPU parallel inference setup, the generation time for a single video is approximately 355 seconds.

OpenSora-Plan [46] We use the v1.3.0 model version. The preset video resolution is 352×640 with a length of 93 frames. Using a 4-GPU parallel inference setup, the generation time for a single video is approximately 408 seconds.

Mochi 1 [71] We execute the process with the decode type set to "tiled full" and utilize a single GPU pipeline, setting the sampling steps to 64. The preset video resolution is 480×848 with a length of 148 frames. The generation time for a single video is approximately 725 seconds.

Wan2.1 [73] We use the T2V-14B model version. The pre-

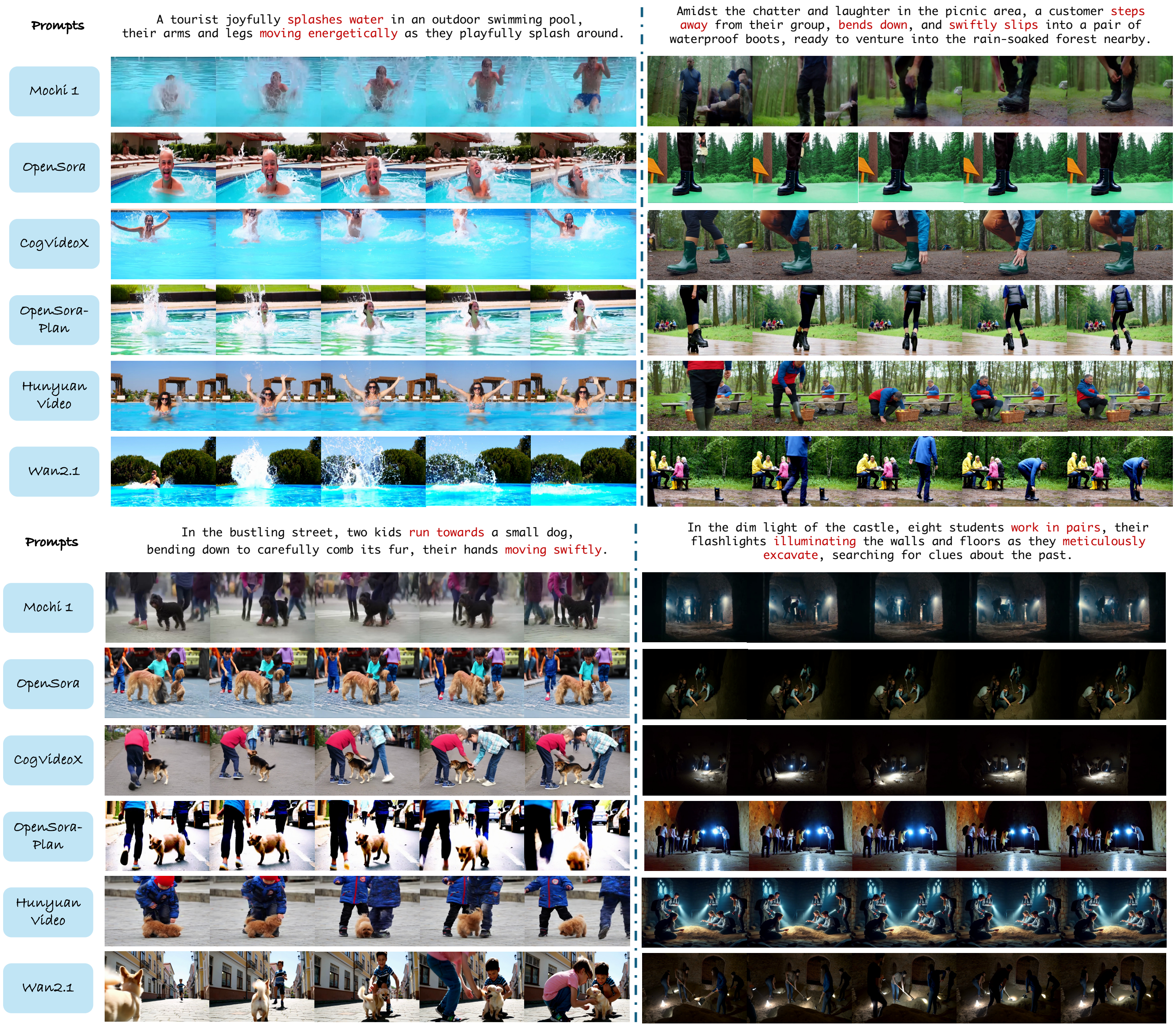


Figure 16. Visualization of Generation Results of Mainstream Models on MMPG-set. Qualitative results on Mochi 1 [71], OpenSora [97], CogVideoX [91], OpenSora-Plan [46], HunyuanVideo [38] and Wan2.1 [73] across six movement modes.

set video resolution is 1280×720 with a length of 81 frames. Using 8 GPUs for parallel inference, the generation time for a single video is approximately 912 seconds.

D.2. Human Annotation

We recruit three annotators and instruct them to score each video based on five previously defined assessment aspects. These aspects are Commonsense Adherence, Motion Smoothness, Object Integrity Score, Perceptible Amplitude, and Temporal Coherence. For each video's motion quality, the annotators assign scores according to the rating criteria outlined. Our annotation process employs a Likert scale [54], with each dimension rated on five levels. Annotators receive detailed descriptions for each dimension to guide their scoring decisions. Our annotation interface is shown in Fig. 15. To ensure a focused evaluation of each aspect, we divide the overall task into five separate annotation packages. In each package, annotators watch the corresponding videos and evaluate only one specific dimension. This approach allows annotators to concentrate on a single aspect of video quality at a time, potentially improving the accuracy and consistency of their assessments. By structuring the annotation process in this way, we aim to obtain more reliable and targeted evaluations for each of the five dimensions of video motion quality.

E. Qualitative Analysis

To identify where current T2V models exhibit limited capabilities, we qualitatively demonstrate the generation results of T2V models. We select 4 challenging prompts from our benchmark, spanning 6 movement modes for video generation. Fig. 16 reveals four critical failure modes: **Object Persistence Paradox:** Models frequently violate object identity continuity during motion. **Structural Degeneration:** Dynamic motion induces catastrophic shape distortions. **Temporal Artifacts:** The generated motion exhibits abrupt discontinuities masked by artificial blurring. **Newtonian Violations:** Fundamental physics laws are systematically broken, particularly in energy conservation.

Upon closer examination of the videos generated by various models, we observe significant disparities in quality and adherence to realistic motion. Mochi 1 [71], OpenSora [97], and OpenSora-Plan [46], for instance, produce videos plagued by severe blurring and artifacts, substantially degrading overall video quality. While CogVideoX [91] and HunyuanVideo [38] demonstrate smoother motion, they struggle with maintaining object integrity, often resulting in unnatural distortions of shape during movement sequences.

Notably, we find that Wan2.1 [73] exhibits the most promising performance among the evaluated models. It generates videos with smooth motion that adhere well to basic physical principles, aligning closely with our fundamental visual expectations. Upon careful observation of task-specific details such as object shapes and limb movements, Wan2.1's outputs appear more natural and consistent. Moreover, it demonstrates a superior ability to accurately represent the amplitude and scale of specific movements as described in the prompts.

These observations underscore the ongoing challenges in text-to-video generation, particularly in maintaining consistency, physical plausibility, and natural motion across diverse scenarios. While progress is evident in some models, there remains significant room for improvement in addressing these critical aspects of video generation.

References

[1] Runway gen3. Accessed February 25, 2025 [Online] https://app.runwayml.com/, 2025. 2

[2] Sora. Accessed February 25, 2025 [Online] https://openai.com/index/sora/, 2025. 1, 2, 4

[3] kling. Accessed February 25, 2025 [Online] https://klingai.com/, 2025. 1, 3

[4] Pika labs. Accessed February 25, 2025 [Online] https://www.pika.art/, 2025. 4, 1

[5] Veo 2. Accessed February 25, 2025 [Online] https://deepmind.google/technologies/veo/veo-2/, 2025. 2, 3

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[7] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2

[8] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1, 6

[9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 7

[10] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 1, 6

[11] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 3

[12] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 1, 2

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 5, 6, 7

[14] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1, 4, 5, 6, 2, 3

[15] 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, 2023. 2, 4, 1

[16] 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, 2024. 2, 4, 1

[17] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[18] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 5

[19] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 7

[20] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 4, 2

[21] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 2, 5, 6

[22] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Irc-gan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019. 2

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[24] Daniel J. Felleman and David C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1 1:1–47, 1991. 4

[25] Xiaokun Feng, Haiming Yu, Meiqi Wu, Shiyu Hu, Jintao Chen, Chen Zhu, Jiahong Wu, Xiangxiang Chu, and Kaiqi Huang. Narrlv: Towards a comprehensive narrative-centric evaluation for long video generation models. *arXiv preprint arXiv:2507.11245*, 2025. 3

[26] Deborah M Gordon. The ecology of collective behavior. *PLoS biology*, 12(3):e1001805, 2014. 1

[27] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33: 16761–16772, 2020. 2

[28] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 1, 3, 5

[29] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. 2

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 3

[31] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 2

[32] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[33] 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

[34] 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. 1, 2, 3, 5, 6, 7

[35] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22747–22757, 2023. 2

[36] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 5, 4, 6

[37] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020. 2

[38] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2, 4, 6, 8, 9, 10

[39] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 4, 6

[40] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020. 4, 6, 7

[41] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024. 3

[42] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[43] 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. 4, 6, 7

[44] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 1

[45] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2025. 1, 6

[46] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 1, 2, 4, 6, 8, 9, 10

[47] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 3

[48] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 4, 1

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4, 5

[50] 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*, 36:62352–62387, 2023. 1

[51] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 1, 2, 3, 5, 7

[52] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 3

[53] Rayeesa Mehmood, Rumaan Bashir, and Kaiser J Giri. Vtmgan: video-text matcher based generative adversarial network for generating videos from textual description. *International Journal of Information Technology*, 16(1):221–236, 2024. 2

[54] Tomoko Nemoto and David Beglar. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, pages 1–6, 2014. 7, 9

[55] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022. 4, 5, 2, 3

[56] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023. 1

[57] TA Palomaki, JD Teufel, RW Simmonds, and Konrad W Lehnert. Entangling mechanical motion with microwave fields. *Science*, 342(6159):710–713, 2013. 1

[58] Marina A Pavlova. Biological motion processing as a hallmark of social cognition. *Cerebral cortex*, 22(5):981–995, 2012. 1

[59] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, et al. Movie gen: A cast of media foundation models, 2025. 2, 5

[60] 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. 1, 2, 5, 6, 7

[61] Ralph H Rasshofer, Martin Spies, and Hans Spies. Influences of weather phenomena on automotive laser radar systems. *Advances in radio science*, 9:49–60, 2011. 1

[62] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 6

[63] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 5, 4

[64] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. *arXiv preprint arXiv:2501.09781*, 2025. 2

[65] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1, 3

[66] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017. 1

[67] 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

[68] Jos Stam. Stable fluids. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 779–786. 2023. 1

[69] 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. 1

[70] Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan.

From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674*, 2024. 1

[71] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024. 1, 2, 6, 8, 9, 10

[72] Qwen Team. Qwen2.5: A party of foundation models, 2024. 5

[73] Wan Team. Wan: Open and advanced large-scale video generative models. 2025. 6, 8, 9, 10

[74] 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. 2, 4, 6, 7

[75] I Telecom. Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. *ITU-R Rec. BT*, 500, 2000. 4

[76] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1, 3

[77] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. 2

[78] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 4, 1

[79] 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*, 2(3):4, 2023. 1

[80] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024. 1, 5, 6

[81] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicvideo-v2: Multi-stage high-aesthetic video generation, 2024. 2

[82] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 2, 6, 7

[83] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6, 7

[84] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2

[85] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 2, 6, 7

[86] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pages 54015–54029. PMLR, 2024. 4, 2

[87] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, 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, 2023. 2

[88] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 4, 3

[89] 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. 2023. 2

[90] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 5, 6

[91] 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. 1, 2, 4, 6, 8, 9, 10

[92] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6, 7

[93] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 2

[94] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005. 2, 6, 7

[95] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 1

[96] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2, 6, 7

[97] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 1, 2, 4, 6, 8, 9, 10

[98] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 5, 6

[99] 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. 2