

VMBench: A Benchmark for Perception-Aligned Video Motion Generation

Xinran Ling^{1*} Chen Zhu^{1*} Meiqi Wu^{1,3*} Hangyu Li¹ Xiaokun Feng^{1,2}
 Cundian Yang¹ Aiming Hao¹ Jiashu Zhu¹ Jiahong Wu^{1†} Xiangxiang Chu¹
¹ AMAP, Alibaba Group ² CRISE, Institute of Automation, Chinese Academy of Sciences
³ School of Computer Science and Technology, University of Chinese Academy of Sciences

Abstract

*Video generation has advanced rapidly, improving evaluation methods, yet assessing video’s motion remains a major challenge. Specifically, there are two key issues: 1) current motion metrics do not fully align with human perceptions; 2) the existing motion prompts are limited. Based on these findings, we introduce **VMBench**—a comprehensive Video Motion Benchmark that has perception-aligned motion metrics and features the most diverse types of motion. VMBench has several appealing properties: (1) **Perception-Driven Motion Evaluation Metrics**, we identify five dimensions based on human perception in motion video assessment and develop fine-grained evaluation metrics, providing deeper insights into models’ strengths and weaknesses in motion quality. (2) **Meta-Guided Motion Prompt Generation**, a structured method that extracts meta-information, generates diverse motion prompts with LLMs, and refines them through human-AI validation, resulting in a multi-level prompt library covering six key dynamic scene dimensions. (3) **Human-Aligned Validation Mechanism**, we provide human preference annotations to validate our benchmarks, with our metrics achieving an average 35.3% improvement in Spearman’s correlation over baseline methods. This is the first time that the quality of motion in videos has been evaluated from the perspective of human perception alignment. Additionally, we release VMBench at <https://github.com/AMAP-ML/VMBench>, setting a new standard for evaluating and advancing motion generation models.*

1. Introduction

Video generation models [2, 3, 37, 43, 55, 78, 90] have advanced significantly, particularly text-to-video (T2V) models [12, 45, 70, 96], which perform well in generating static content but struggle to capture motion in alignment with hu-

man perception. To enhance dynamic motion generation, it is crucial to establish perception-aligned motion evaluation.

However, previous methods [29, 49, 59, 64, 75] primarily assess the static content of videos, with dynamic evaluation largely limited to motion smoothness [33, 44, 50]. This makes it challenging to capture motion deficiencies [69, 94], including but not limited to spatiotemporal inconsistencies and violations of physical laws. *Therefore, developing a more comprehensive evaluation metric that better aligns with human perception is crucial.*

To address this, we introduce Video Motion Benchmark (VMBench) (Fig. 1), the first benchmark focused on comprehensive video motion assessment. It features the most diverse range of motion types, covering 969 categories—more than any other benchmark [27, 33, 44, 49, 50, 68] (see Fig. 12 in the Appendix). Additionally, it integrates perception-aligned motion metrics to systematically evaluate the motion generation quality of video generation models.

First, we introduce Perception-Driven Motion Evaluation Metrics (PMM), the first evaluation framework explicitly designed to align with human perception of motion quality. PMM comprises five key components: Object Integrity Score (OIS), Perceptible Amplitude Score (PAS), Temporal Coherence Score (TCS), Motion Smoothness Score (MSS), and Commonsense Adherence Score (CAS). Unlike previous methods that primarily focus on motion smoothness [29, 59, 64, 75], PMM provides a more comprehensive assessment by evaluating motion quality in terms of spatiotemporal inconsistencies and violations of physical laws.

Next, we propose Meta-Guided Motion Prompt Generation (MMPG), an expandable framework that encompasses the most comprehensive range of motion types, grounded in physics [56, 60, 65, 67] and cognitive science [25, 57]. The evaluation system covers six movement modes: fluid dynamics, biological motion, mechanical motion, weather phenomena, collective behavior, and energy transfer (Fig. 1). Specifically, MMPG first extracts subjects, places, and actions from VidProm [79], Place365 [97], and

*Equal Contribution

†Corresponding authors

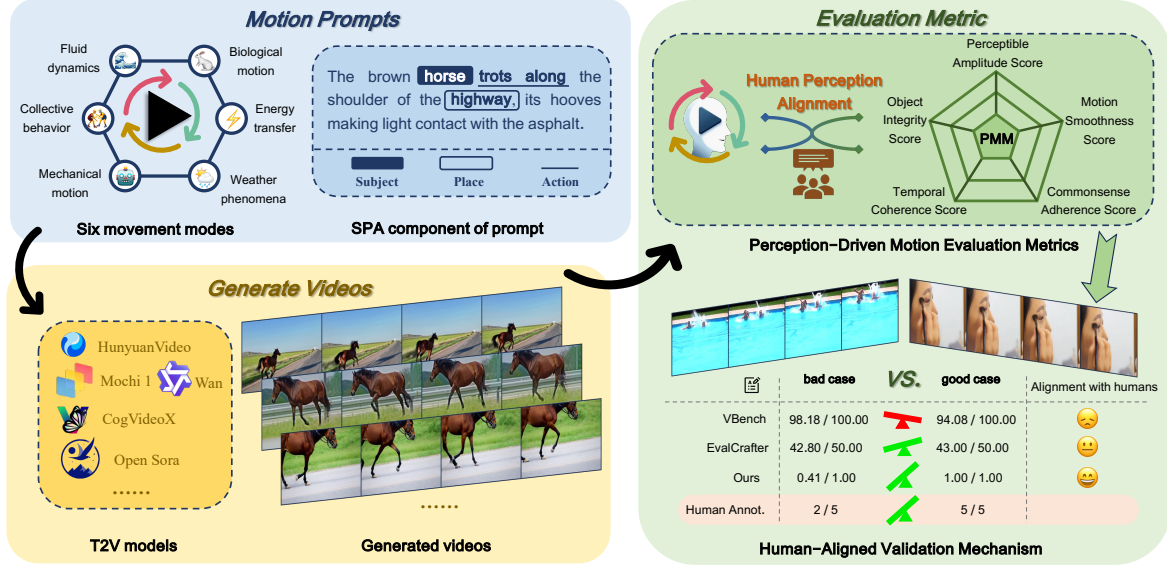


Figure 1. Overview of **VMBench**. Our benchmark encompasses six principal categories of motion patterns, with each prompt constructed as a comprehensive motion structured around three core components: subject, place, and action. We propose a novel multi-dimensional video motion evaluation comprising five human-centric quality metrics derived from perceptual preferences. Utilizing videos generated by popular T2V models, we conduct systematic human evaluations to validate the effectiveness of our metrics in capturing human perceptual preferences.

other datasets [8, 10, 14, 89], then leverages LLMs to enrich metadata and generate detailed motion descriptions. We validate the prompts through human-AI collaboration with DeepSeek-R1 [21] to ensure coherence and rationality. To reduce costs, we curated 1,050 user prompts and use T2V models to generate motion videos for evaluation.

Finally, we provide human preference annotations to validate our benchmarks, showing that our metrics achieve an average 35.3% improvement in Spearman’s correlation [93] over current methods [9, 13, 33, 50, 59, 73, 81, 84, 95]. This result highlights the enhanced alignment of our evaluation metrics with human perception. We will open-source VMBench, including all prompts, evaluation methods, generated videos, and human preference annotations, and also include more video generation models in VMBench to drive forward the field of video motion generation. Our contributions can be summarized as follows:

- **First-Ever Human Perception-Aligned Motion Evaluation:** We pioneer the evaluation of motion quality in videos from the perspective of human perception alignment.
- **Meta-Guided Motion Prompt Generation:** A curated prompt set designed to assess diverse motion aspects in video models.
- **Human-Aligned Validation Mechanism:** Our metrics achieve an average 35.3% improvement in Spearman’s correlation over current methods.

2. Related work

2.1. Video Generative Models

With the rapid advancement of AI generative techniques—GANs [22, 26, 36, 41, 52], VAEs [41, 83], autoregressive models [17, 63, 92], and diffusion models [7, 31, 32, 34, 76, 88]—the field of T2V generation has seen remarkable progress. However, while current T2V models [16, 70, 90, 96], excel in generating static visual content, their dynamic performance still deviates significantly from human perception.

Early T2V models [15, 30, 31, 66] have built upon Text-to-Image(T2I) techniques, enhancing their temporal capabilities to generate videos. Then, some works [28, 80, 98] take this further by conducting the diffusion process within the latent space and incorporate the spatio-temporal module [16, 76, 86] within the denoising network, which greatly improves the quality of static content in videos. CogVideo [32, 90] and other closed source works [1, 5, 12, 58] represented by Sora [2] focus on high-quality data filtering on a large-scale dataset, significantly boosting model performance. OpenSora [96] and OpenSora-Plan [45] aim to showcase and open source the remarkable performance of closed-source models. Mochi 1 [70] and HunyuanVideo [37] have improved the dynamic quality of generated videos through new technologies such as asymmetric architecture and full attention mechanism. Beyond optimization of models and datasets, recent stud-

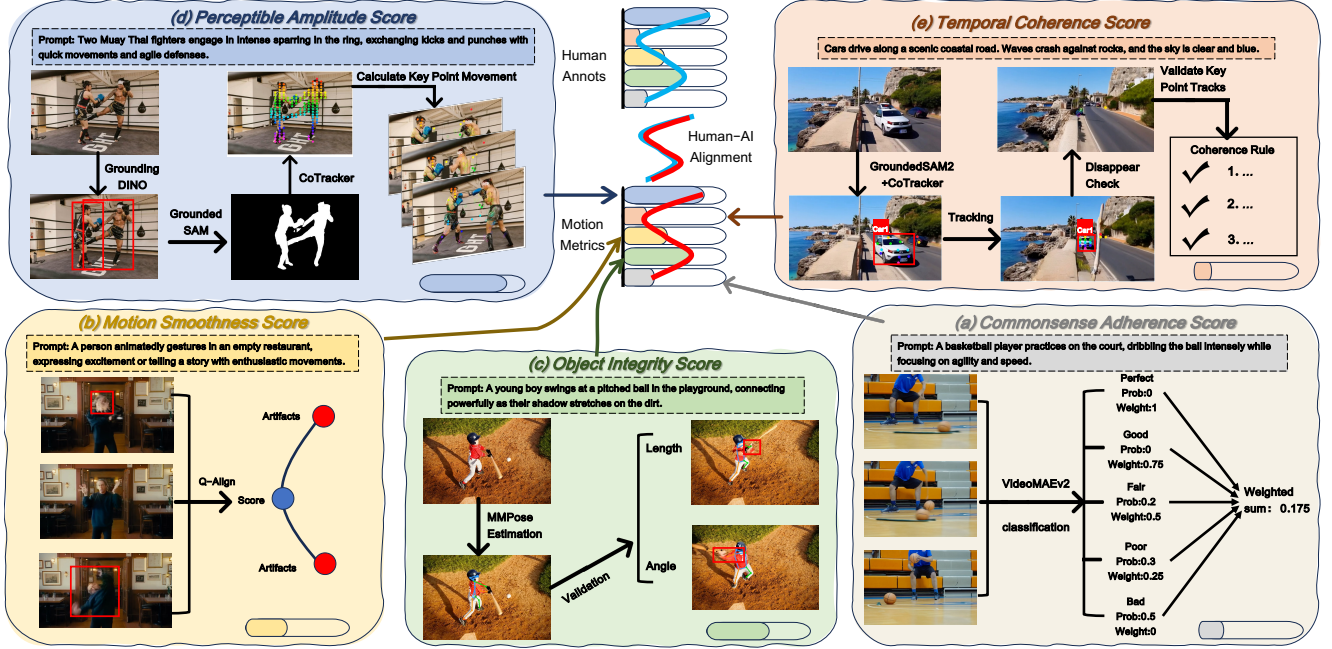


Figure 2. Framework of our Perception-Driven Motion Metrics (PMM). PMM comprises multiple evaluation metrics: Commonsense Adherence Score (CAS), Motion Smoothness Score (MSS), Object Integrity Score (OIS), Perceptible Amplitude Score (PAS), and Temporal Coherence Score (TCS). (a-e): Computational flowcharts for each metric. The scores produced by PMM show variation trends consistent with human assessments, indicating strong alignment with human perception.

ies [3, 5, 40, 46, 51] are beginning to explore the influence of motion on video generation. Specifically, Ma et al. [51] focuses on the motion quality of videos during data selection, in order to enable models to learn motion dynamics even at low resolution training stages. Meanwhile, some works [40, 46] improve the generative capacity of models by incorporating motion information as guidance during the sampling process.

Although these models demonstrate commendable performance across various metrics, the generated videos still exhibit several qualitative motion deficiencies, such as the abrupt disappearance of moving subjects, body distortions, and physically implausible actions, which are difficult to assess using existing metrics.

2.2. Evaluations Metrics for Video Generation

Existing approaches for video motion evaluation can be broadly categorized into three paradigms: 1) feature-based metrics leveraging pre-trained video representations, 2) rule-based metrics employing manually designed scoring mechanisms, and 3) MLLM-based assessments that fine-tune large multimodal language models on human-annotated video quality datasets for perceptual scoring.

Previous feature-based metrics demonstrate limited effectiveness in assessing motion quality. Image-centric metrics like IS [64] and FID [29] inherently ignore tempo-

ral coherence, while FVD [75] extends image-level metrics to videos by modeling spatiotemporal features, it reduces motion dynamics to simplistic distributions. These methods conflate feature-space distances with perceptual motion quality, overlooking artifacts like implausible movement.

Recent benchmarks [24, 33, 50] attempt to conduct systematic evaluations using manually curated metrics. For instance, VBench [33] introduces motion-related criteria such as motion smoothness, quantified via a video frame interpolation model. However, such rule-driven metrics face two inherent limitations in motion assessment: 1) Subjectivity in metric design—predefined rules poorly generalize to diverse motion patterns beyond designer priors; 2) Fragmentary analysis—isolated measurements of individual attributes fail to capture holistic motion quality. Consequently, these heuristics overlook complex phenomena like inertia preservation or multi-object interaction coherence.

Emerging methods [11, 27] employ multimodal large language models (MLLMs) fine-tuned on human preference data to predict perceptual video quality scores. VideoScore [27] aligns scoring with human judgments through preference learning, while VideoPhy [11] specifically targets physical plausibility via physics-focused annotations. Despite improved alignment with subjective evaluations, these approaches exhibit critical limitations for motion assessment: 1) Oversimplified scoring granularity – scalar outputs

obscure specific motion defects; 2) Annotation bias amplification – training data biases toward conspicuous artifacts neglect subtle kinematic violations; 3) Contextual rigidity – domain-specific tuning limits generalization to diverse motion types. While advancing automated evaluation, they struggle to disentangle motion quality from global preference impressions.

Current evaluation paradigms remain fundamentally misaligned with human perception of motion quality due to the above limitations. To bridge this gap, we introduce 5 distinct human-aligned evaluation metrics that systematically quantify motion quality.

3. VMBench

We introduce VMBench, a comprehensive Video Motion Benchmark that includes perception-aligned motion metrics and features the most diverse range of motion types.

3.1. Perception-Driven Motion Evaluation Metrics

When observing videos, the human brain first constructs a holistic understanding of the scene based on prior experiences and physical laws [23], before selectively attending to moving objects to assess motion smoothness and temporal consistency, particularly when objects are momentarily occluded. While rapid movement can effectively capture attention, excessive deviations from typical motion patterns may induce visual disorientation. As illustrated in Fig. 5 in the Appendix, our motion evaluation metrics are designed to emulate this hierarchical perceptual process, transitioning from global scene comprehension to localized motion analysis to ensure coherence and logical consistency.

Commonsense Adherence Score (CAS) is used to evaluate whether generated videos align with human commonsense. As shown in Fig. 2 (a), we train a specialized model capable of assessing the commonsense quality of video content, categorizing it into five levels: Bad, Poor, Fair, Good, and Perfect. First, we collect a dataset comprising 10k generated videos, covering both legacy approaches and popular models [2–4, 15, 16, 37, 45, 90, 96]. Second, we establish perceptual ground truth through systematic pairwise comparisons using VideoReward [47], a reward model trained on 182k human-annotated video preference pairs. This process generates normalized preference scores reflecting human perceptual consensus. Third, we discretize the human preference into five cognitive labels. Using these preference labels, we develop a video classification model to serve as the final CAS predictor, utilizing VideoMAEv2 [77], a spatiotemporal vision transformer. Compared to MLLMs, this architecture demonstrates temporal modeling capabilities through dense frame sampling and 3D convolution operations. We compute the CAS using a Mean Opinion Score (MOS) [74, 85], where predicted probabilities for each class are weighted by their corresponding quality co-

efficients: $CAS = \sum_{i=1}^5 p_i G(i)$, where p_i denotes the predicted probability for the i -th class, and G_i represents the mapping function that converts category to quality weights: $G : i \rightarrow w$ (e.g., mapping class 3 to 0.5). For more details, please refer to our Appendix B.1.

Motion Smoothness Score (MSS) aims to detect unsmoothness, like low-level temporal artifacts and high-level motion blur. It addresses the limitations of prior metrics that suffered from either low-level optical flow bias [38, 39] or oversimplified motion modeling [42], both leading to misalignment with human perception. Considering that natural motion patterns inherently maintain consistent visual quality across temporal sequences, we leverage Q-Align’s [85] aesthetic score to detect artifacts, as illustrated in Fig. 2 (b). An artifact frame is confirmed when a significant drop in Q-Align scores between consecutive frames exceeds a predefined threshold, mimicking human perception of sustained visual anomalies. Further, our approach employs a scene-aware adaptive thresholding mechanism derived from the statistical modeling of real video segments [14, 54] across diverse motion patterns. The final MSS is computed as: $MSS = 1 - \frac{1}{T} \sum_{t=2}^T \mathbb{I}(\Delta Q_t > \tau_s(t))$ where ΔQ_t is the frame-to-frame visual quality degradation magnitude, $\tau_s(t)$ indicates the adaptive threshold and $\mathbb{I}$ is an indicator function. More details are provided in Appendix B.2.

Object Integrity Score (OIS) detects implausible deformations through spatiotemporal analysis of object integrity. Our approach addresses limitations found in previous methods [13], which primarily focus on object-level semantic consistency. Instead, we focus on detecting perceptual issues (e.g., distorted shapes) that are readily noticeable to the human visual system. As illustrated in Fig. 2 (c), we first utilize the MMPose toolkit [20] to detect key points of the primary subjects in the generated videos, which are then used to estimate the subjects’ shapes in each frame. Then, we quantify shape distortions by analyzing whether object shapes violate real-world anatomical constraints. We establish tolerance thresholds for the above constraints through the statistical analysis of natural motion samples [14, 54, 87]. Anatomical violations across consecutive frames will be detected; see more details about anatomical constraints in our appendix B.3. The OIS quantifies structural consistency through: $OIS = \frac{1}{F \cdot K} \sum_{f=1}^F \sum_{k=1}^K \mathbb{I}(\mathcal{D}_f^{(k)} \leq \tau^{(k)})$, in which F and K denote total frame count and anatomical components, respectively. $\mathcal{D}_f^{(k)}$ represents the compound anatomical deviation for component k in frame f , and $\tau^{(k)}$ indicates statistical thresholds.

Perceptible Amplitude Score (PAS) estimates subject motion by separating it from camera motion, unlike traditional motion magnitude estimation methods [73], which often overestimate overall motion due to the influence of camera movement. As presented in Fig. 2 (d), PAS begins with semantic anchoring via GroundingDINO [48] to lo-

calize moving subjects and distinguish them from passively moving elements. Next, GroundedSAM [62] ensures temporally stable subject masks, enabling CoTracker [35] to track key points with high precision. The motion magnitude is then computed based on the average displacement of these key points. Furthermore, PAS accounts for the context-dependent nature of human motion perception, adapting thresholds across different scenarios. To quantify this variability, we derive a set of perceptual motion magnitude thresholds for various scenarios through the statistical analysis of existing video datasets [14, 54]. These thresholds serve as the foundation for computing a motion score for each video. The PAS is computed as $PAS = \frac{1}{T} \sum_{t=1}^T \min\left(\frac{\bar{D}_t}{\tau_s}, 1\right)$, where $\bar{D}_t$ denotes frame-level motion amplitude, computed as the average displacement of tracked key points for active subjects in frame t and τ_s is the perceptual motion threshold for scenario s . More Details can be found in appendix B.4.

Temporal Coherence Score (TCS) detects the temporal consistency of object motion by analyzing frame-to-frame transitions, addressing the limitation of existing metrics [13, 59] that struggle to differentiate between natural movement and abrupt, unrealistic changes. We combine video object tracking with rule-based validation to detect abnormal disappearance/reappearance patterns, thereby distinguishing natural progression from temporal discontinuities. As shown in Fig. 2 (e), first GroundedSAM2 [61] performs 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 the CoTracker [35]. It tracks dense key points on target objects and constructs their motion trajectories. We then analyze their 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 implemented via trajectory analysis. These rules account for common scenarios, including 1) objects reappearing after occlusion or disappearing behind obstacles, 2) objects entering or exiting frame boundaries, and 3) apparent size changes due to depth perception, such as objects appearing larger when moving closer or smaller when moving farther away. Through this filtering, we systematically exclude normal events. The TCS quantifies anomalous discontinuities via $TCS = 1 - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\mathcal{A}_i \wedge \neg \mathcal{R})$, where N is the total object instances, $\mathcal{A}_i$ indicates the existence status of the i -th object, and $\mathcal{R}$ validates legitimate transitions (see Appendix B.5).

3.2. Meta-Guided Motion Prompt Generation

The existing benchmarks [27, 33, 50] are constrained by limited and simplistic motion types (As shown in Fig. 12

of the Appendix), making them inadequate for comprehensively evaluating the motion generation capabilities of video generation models. To address this limitation, we propose an extensible framework—Meta-Guided Motion Prompt Generation (MMPG)—that generates prompts capturing complex motion patterns, as illustrated in Fig. 3. Built upon MMPG, we introduce VMBench, a benchmark specifically designed for rigorous evaluation of complex motion generation. We find that VMBench provides the most comprehensive coverage of motion types and the most detailed prompt descriptions, making it an effective benchmark for evaluating the dynamic motion generation capabilities of video generation models. In this section, we elaborate on the process of prompt generation and refinement. More details are available in Appendix C.1.

Meta-information Extraction. We decompose motion descriptions into three key metadata elements: Subject (S), Location (P) and Action (A). We employ Qwen-2.5 [71] to extract the metadata library $\{S, P, A\}$ from existing video-text pairs [18, 58, 79, 89]. To ensure novel prompts and enhance textual diversity, we expand the metadata in three key aspects. For subject (S) classification, we categorize subject types into human, animal, and object, curate a list of principal nouns identifiable by [48, 62], and use GPT-4o [6] to generate descriptions with varying entity counts (1, 2, n). For place (P) descriptions, we incorporate data from [97], filtering out redundant information. For action (A) expansion, we sample human actions from [14] and use LLM to extend possible actions for animals and objects.

Self-Refining Prompt Generation. We randomly sample elements from the metadata library $\{S, P, A\}$ to form metadata sets (S^i, P^j, A^k) and use GPT-4o to assess their semantic coherence and logical consistency. For coherent sets, we generate prompts of 10 to 60 words that describe the corresponding motion scenarios. To ensure accuracy, we introduce a verification mechanism inspired by [50], where the generated prompts are re-evaluated by GPT-4o to verify their consistency with the metadata elements. Through an iterative refinement and filtering process, we generate approximately 50,000 high-quality prompt candidates. However, grammatical correctness alone is insufficient: prompts must also align with physical reality to effectively assess video content. Therefore, we also eliminate descriptions of unrealistic or physically implausible scenarios to ensure the feasibility and authenticity of the prompts.

Human-LLM Joint Validation. To reduce manual labor and enhance evaluation effectiveness, we further examine and filter prompts through a Human-LLM Joint Validation process. This process combines the strengths of both automated reasoning and human judgment to ensure high-quality motion descriptions. We first leverage the powerful reasoning capabilities of Deepseek R1 [21] to evaluate whether each prompt describes a reasonably realistic mo-

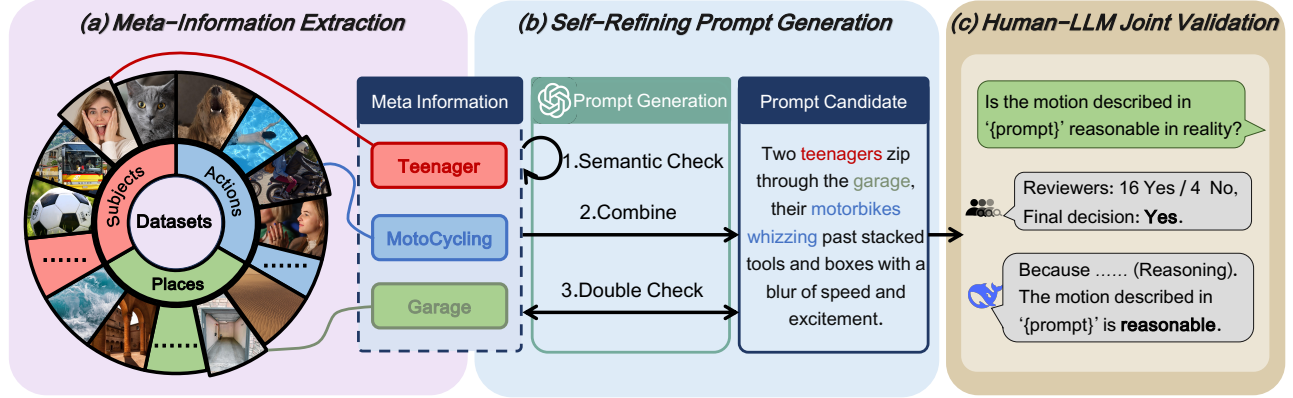


Figure 3. Framework of our **Meta-Guided Motion Prompt Generation (MMPG)**. MMPG consists of three stages: **(a) Meta-information Extraction**: Extracting Subjects, Places, and Actions from datasets such as VidProm [79], Didemo [8], MSR-VTT [89], WebVid [10], Place365 [97], and Kinect-700 [14]. **(b) Self-Refining Prompt Generation**: Generating and iteratively refining prompts based on the extracted information. **(c) Human-LLM Joint Validation**: Validating the prompts through a collaborative process between humans and DeepSeek-R1 to ensure their rationality.

tion. After filtering out implausible descriptions, we curate a diverse set of 1,050 prompts with varied metadata. Examples are available in Appendix C.2.

4. Experiments

4.1. Experimental Setup

Implementation Details. Our benchmark evaluates six popular text-to-video models: OpenSora [96], CogVideoX [90], OpenSora-Plan [45], Mochi 1 [70], HunyuanVideo [37], and Wan2.1 [72]. To provide a richer variety of motion types, we develop the MMPG-set, which comprises 1,050 prompts based on MMPG across six movement modes to assess performance in motion generation. Each model generates 1,050 videos corresponding to the MMPG-set, resulting in a total of 6,300 videos. To ensure a fair comparison, we maintain the hyperparameter settings as defined in each model’s project demo. For each prompt, we generate one video for evaluation using only the initial seed. The inference process is executed using 8 Nvidia H20 GPUs. Details of the inference process for each model can be found in Appendix D.1. We then randomly sample 200 videos from each model’s output, resulting in a total of 1,200 videos for human-aligned validation experiments.

Comparison Metrics. The current comparison approaches fall into two categories: Rule-based metrics and Multimodal Large Language Model (MLLM) prompting. (1) Rule-based metrics assess four dimensions, including Perceptible Amplitude, which evaluates through RAFT optical flow magnitude analysis combined with structural motion consistency via 4 frame SSIM averaging [73, 82], following established protocols [33]. Temporal Coherence is measured using DINO [13] and CLIP [59] feature tracking, which relies on cosine similarity between consecu-

tive frames. Motion Smoothness is assessed through a hybrid method that combines interpolation error [42] with Dover’s video quality assessment [84], while Object Integrity is evaluated via dual validation using optical flow warping error [38, 39] and semantic consistency checks. (2) The MLLM evaluation includes five cutting-edge models: LLaVA-NEXT-Video [95], MiniCPM-V-2.6 [91], InternVL2.5 [19], Qwen2.5-VL [9], and InternVideo2.5 [81]. These models are evaluated through a standardized assessment process, which involves processing uniformly sampled video frames at 2 frames per second to maintain essential motion patterns while managing computational limits. The evaluation dimensions encompass five aspects: amplitude, coherence, integrity, smoothness, and common-sense adherence, with scoring for each dimension using a 1–5 scale. In particular, to ensure fair comparison, consistent frame sequences and evaluation criteria are maintained across all models.

Metrics. 1) *Spearman correlation* [93]: Spearman’s rank correlation coefficient, often denoted by ρ , is a measure of the strength and direction of the association between two ranked variables. It assesses how well the relationship between two variables can be described by a monotonic function. Spearman correlation is nonparametric and can effectively evaluate associations in datasets where variables may not follow a normal distribution. Unlike Pearson correlation, which captures linear relationships, Spearman correlation focuses on rank-based associations, making it more robust to outliers and suitable for ordinal data or scenarios with nonlinear dependencies. 2) *Accuracy*: To validate how accurately our metrics align with human preferences, we conduct pairwise comparisons across the 1,200 annotated videos (200 prompts $\times$ 6 models). For each prompt, we

Method	Avg. $\rho(\uparrow)$	CAS $\rho(\uparrow)$	MSS $\rho(\uparrow)$	OIS $\rho(\uparrow)$	PAS $\rho(\uparrow)$	TCS $\rho(\uparrow)$
<i>Rule-based</i>						
SSIM [82]	1.6	-0.9	-12.1	8.3	17.8	-4.8
RAFT* [73]	-1.7	-0.7	-17.0	-16.6	47.7	-21.9
CLIP* [59]	15.0	21.5	36.5	31.7	-42.7	28.0
DINO* [13]	9.4	11.0	27.2	27.4	-45.8	27.4
Dover Technical [†] [84]	20.6	40.2	32.6	34.5	-6.2	2.2
AMT* [33]	3.6	2.8	18.1	19.6	-40.4	17.8
Warping Error [†] [50]	-6.0	-11.4	-19.1	-9.0	33.0	-23.6
<i>MLLM Prompting</i>						
LLaVa-NEXT-Video [95]	17.1	15.2	11.0	16.1	30.8	12.4
MiniCPM-V-2.6 [91]	16.2	6.8	17.5	-2.2	27.3	31.6
IntenVL2.5 [19]	19.3	18.7	20.5	14.1	42.5	1.1
Qwen2.5-VL [9]	15.3	24.7	19.8	3.7	14.2	14.0
InternVideo2.5 [81]	26.9	22.7	21.9	29.6	44.3	15.8
<i>Model-Rule Synergy</i>						
PPM (Ours)	62.2	69.9	77.1	65.8	65.2	54.5

Table 1. Correlation Analysis Between Evaluation Metrics and Human Scores via Spearman Correlation Coefficient ($\rho \times 100$). Superscript * and [†] denote implementations following VBench [33] and EvalCrafter [50] respectively. Yellow backgrounds in Rule-based indicate specific dimension baseline.

evaluate all 15 possible video pairs generated by different models (total 3,000 pairs). Human preference labels are determined by comparing the average expert ratings across all five dimensions (OIS, MSS, CAS, TCS, PAS) – the video with higher aggregated human scores is designated as the “ground truth” preferred sample. Concurrently, we compute metric preferences by comparing videos’ aggregated PMM scores under identical criteria. Alignment accuracy is quantified as the percentage of pairs where PMM preferences match human judgments, with ties excluded to prioritize unambiguous decisions.

4.2. Comparison with State-of-the-Art

Human-aligned Validation Mechanism. We invite three domain experts to independently annotate each sample based on PMM (Perceptible Amplitude, Temporal Coherence, Object Integrity, Motion Smoothness, and Commonsense Adherence), resulting in 6,000 detailed ratings with high interannotator agreement. Finally, we obtain Likert scores [53] from humans for each video across the five dimensions included in PMM. More details can be found in Appendix D.2. We assess the alignment between metrics and human perception by calculating the Spearman correlation [93] between metric scores and expert ratings. A higher Spearman correlation signifies a stronger alignment with human perception.

Comparison with Alternative Metrics. As shown in Table 1, in the MSS evaluation, even advanced metrics

CAS	MSS	OIS	PAS	TCS	Accuracy (%) $\uparrow$
<i>Removal-based ablation</i>					
✓	✓	✓	✓	✓	70.6
✓	✓	✓	✓	✗	66.9
✓	✓	✓	✗	✓	68.7
✓	✓	✗	✓	✓	64.6
✓	✗	✓	✓	✓	65.2
✗	✓	✓	✓	✓	64.1
<i>Addition-based ablation</i>					
✗	✗	✗	✗	✗	-
✓	✗	✗	✗	✗	58.9
✓	✓	✗	✗	✗	66.1
✓	✓	✓	✗	✗	67.3

Table 2. Ablation Study on our Motion Metrics. Prediction accuracy (%) of different metric combinations against human preferences is calculated. *Removal-based ablation* show the effect of individually ablating each metric, while in the *addition-based ablation*, we progressively add each metric to observe its impact.

such as AMT [42] (18.1%) and Warping Error [39] (-19.1%) demonstrate limited discriminative power and produce counterintuitive results, particularly in the context of complex deformations. The OIS evaluation also highlights similar shortcomings, with DINO [13] at 27.4% and Dover [84] at 34.5%, both of which fail to capture human sensitivity to structural preservation during motion. In relation to the evaluation of PAS, the SSIM [82] and RAFT [73] metric based on rules show an alignment efficacy of only 17.8% and 47.7%, respectively. In contrast, our approach achieves a remarkable 65.2% alignment efficacy. For the TCS evaluation, the rule-based metrics CLIP [59] and DINO [13] achieve only 28.0% and 27.4% alignment efficacy, respectively, and fail to capture human tolerance for minor inconsistencies while maintaining physical plausibility. In comparison, our method achieves an impressive 54.5% alignment efficacy. VBench [33] includes RAFT, CLIP, DINO, and AMT, while EvalCrafter [50] incorporates Dover Technical and Warping Error. According to the data presented in the table, *overall, compared to ours, the motion evaluations of VBench [33] and EvalCrafter [50] reveal significantly lower correlations with human perception*. MLLMs exhibit partial competence in evaluations of the Physical Adequacy Score (PAS) (e.g., InternVideo2.5: 44.3%) *Despite their general capabilities, MLLMs average only a correlation of 10.0%–30.0% in all dimensions, highlighting a fundamental misalignment with human perception in evaluating motion quality*.

4.3. Ablation Study

Ablation Study of Motion Metrics. According to Table 2, removing any single metric from the evaluation of the

basic variants results in a significant decline in accuracy, which highlights the importance of each dimension’s evaluation metric within the overall framework. Notably, the removal of the CAS metric leads to the most pronounced performance drop, with accuracy falling to 64.1%, exceeding the impact of other single-dimension ablations. This demonstrates the critical role of the CAS metric in assessing video quality, aligning closely with the key factors prioritized by humans when perceiving video quality. For the performance-oriented variants, we emulate the human perceptual information processing flow by gradually adding evaluation metrics (refer to Fig. 5 in the Appendix), with each additional metric significantly enhancing overall accuracy. This not only demonstrates the effectiveness of incrementally increasing evaluation dimensions but also further verifies the consistency of our evaluation approach with human perceptual mechanisms.

4.4. Qualitative Analysis

Alignment of PMM with Human Perception. As shown in Fig. 4, the correlation between human scores across these five evaluation dimensions and our evaluation metrics in the same five dimensions remains consistent. For example, both demonstrate a strong correlation between OIS, CAS, and MSS, as well as a weak correlation between PAS and the other metrics. Specifically, as shown in Fig. 4 (a), PAS has a negative correlation with other dimensions, such as the Object Integrity Score (OIS), where $\rho = -0.18$. This might be because high dynamic amplitudes in videos cause distortions and artifacts, which lower structural and temporal coherence scores. Conversely, OIS shows strong positive correlations with MSS ($\rho = 0.59$) and CAS ($\rho = 0.50$), suggesting that it reflects physical plausibility and motion rationality well. TCS shows low correlation with other dimensions, which can provide a more comprehensive assessment perspective. PAS and structural/temporal metrics exhibit a negative correlation, which challenges conventional optical-flow-based evaluation frameworks. Additionally, the isolation of PAS highlights the importance of separately disentangling motion magnitude in motion video evaluation. Fig. 4 (b) shows the correlations among the proposed evaluation metrics, which align with human perception.

Assessing Video Generation Models with PMM. As shown in Table 3, we evaluate several leading video generation models using our PMM metrics, including Mochi 1[70], OpenSora [96], CogVideoX [90], OpenSora-Plan [45], HunyuanVideo [37], and Wan2.1 [72]. Our findings indicate that Wan2.1 delivers the best performance in generating motion videos, producing results that appear more realistic. For a clearer visual comparison of the dynamic videos generated by each model, please refer to Fig. 16 in the Appendix.

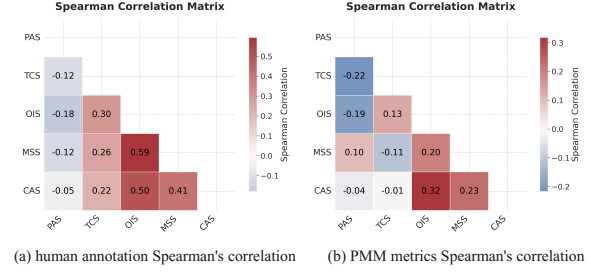


Figure 4. Correlation Matrix Analysis of Metrics Within Different Evaluation Mechanisms. (a): Spearman Correlation Matrices for human annotations; (b): Spearman Correlation Matrices for our PMM metrics.

Models	Avg.(%) ↑	CAS	MSS	OIS	PAS	TCS
Mochi 1[70]	53.2	37.7	62.0	68.6	14.4	83.6
OpenSora [96]	51.6	31.2	61.9	73.0	3.4	88.5
CogVideoX [90]	60.6	50.6	61.6	75.4	24.6	91.0
OpenSora-Plan [45]	58.9	39.3	76.0	78.6	6.0	94.7
HunyuanVideo [37]	63.4	51.9	81.6	65.8	26.1	96.3
Wan2.1 [72]	78.4	62.8	84.2	66.0	17.9	97.8

Table 3. Performance of Video Generation Models on VMBench. We evaluated six open-source video generate models [37, 45, 70, 72, 90, 96] using VMBench. A higher score indicates better performance for a category.

5. Conclusion

Video motion authenticity remains a critical challenge in the creation of generate content. In response, we introduce VMBench, the first open-source benchmark for motion quality evaluation, integrating motion metrics with human-aligned assessment to reveal deficiencies in existing models’ ability to generate physically plausible movements. To support the research community, we offer three essential resources: 1) a standardized framework for detecting motion artifacts overlooked by traditional metrics; 2) actionable diagnostic tools aimed at guiding model optimization; and 3) quantitative standards that align technical advancements with human perceptual expectations. By establishing a common framework for development in this field, VMBench enables systematic tracking of motion generation, encourages targeted model improvements, and ultimately supports the creation of videos that achieve both visual fidelity and dynamic realism. However, while the benchmark’s evaluation metrics are aligned with general human perception, they may not fully capture subtle differences in perception that stem from individual viewer experiences and preferences. Addressing these limitations presents opportunities for further refinement, aiming to ensure a more comprehensive and adaptable approach to video quality evaluation.

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
- [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. 2, 6
- [9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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. 2, 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. 2, 4, 5, 6
- [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
- [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
- [17] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training 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
- [21] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 2, 5
- [22] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Ircgan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019. 2
- [23] Daniel J. Felleman and David C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1 1:1–47, 1991. 4
- [24] 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
- [25] Deborah M Gordon. The ecology of collective behavior. *PLoS biology*, 12(3):e1001805, 2014. 1
- [26] Shir Gur, Sagie Benaïm, 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
- [27] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang 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
- [28] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. 2
- [29] 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

- [30] 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
- [31] 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
- [32] 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
- [33] Ziqi Huang, Yanan 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
- [34] 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
- [35] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 5
- [36] 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
- [37] 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
- [38] 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
- [39] 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
- [40] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhui 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
- [41] 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
- [42] 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
- [43] Zhengqi Li, Richard Tucker, Noah Snively, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 1
- [44] 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
- [45] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Lihuan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 1, 2, 4, 6, 8
- [46] 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
- [47] 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
- [48] 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
- [49] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Shihuo 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
- [50] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong 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
- [51] 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
- [52] Rayeesa Mehmood, Rumaan Bashir, and Kaiser J Giri. Vtm-gan: video-text matcher based generative adversarial network for generating videos from textual description. *International Journal of Information Technology*, 16(1):221–236, 2024. 2
- [53] Tomoko Nemoto and David Beglar. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, pages 1–6, 2014. 7
- [54] 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 *Pro-*

- ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022. 4, 5
- [55] 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
- [56] TA Palomaki, JD Teufel, RW Simmonds, and Konrad W Lehnert. Entangling mechanical motion with microwave fields. *Science*, 342(6159):710–713, 2013. 1
- [57] Marina A Pavlova. Biological motion processing as a hallmark of social cognition. *Cerebral cortex*, 22(5):981–995, 2012. 1
- [58] 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
- [59] 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
- [60] 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
- [61] 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
- [62] 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
- [63] 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
- [64] 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
- [65] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017. 1
- [66] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyan 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
- [67] Jos Stam. Stable fluids. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 779–786. 2023. 1
- [68] 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
- [69] 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
- [70] Genmo Team. Mochi 1. <https://github.com/genmoai/models>, 2024. 1, 2, 6, 8
- [71] Qwen Team. Qwen2.5: A party of foundation models, 2024. 5
- [72] Wan Team. Wan: Open and advanced large-scale video generative models. 2025. 6, 8
- [73] 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
- [74] I Telecom. Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. *ITU-R Rec. BT*, 500, 2000. 4
- [75] 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
- [76] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscape text-to-video technical report, 2023. 2
- [77] 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
- [78] 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
- [79] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024. 1, 5, 6
- [80] 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
- [81] 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
- [82] 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
- [83] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Gen-

- erating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2
- [84] 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
- [85] 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 llms for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pages 54015–54029. PMLR, 2024. 4
- [86] 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
- [87] 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
- [88] 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
- [89] 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. 2, 5, 6
- [90] 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
- [91] 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
- [92] 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
- [93] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005. 2, 6, 7
- [94] 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
- [95] 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
- [96] 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
- [97] 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
- [98] 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