

PAI-Bench: A Comprehensive Benchmark For Physical AI

Fengzhe Zhou^{1,*} Jiannan Huang^{1,*} Jialuo Li^{1,*} Deva Ramanan² Humphrey Shi¹
¹Georgia Tech ²CMU
 *Equal contribution

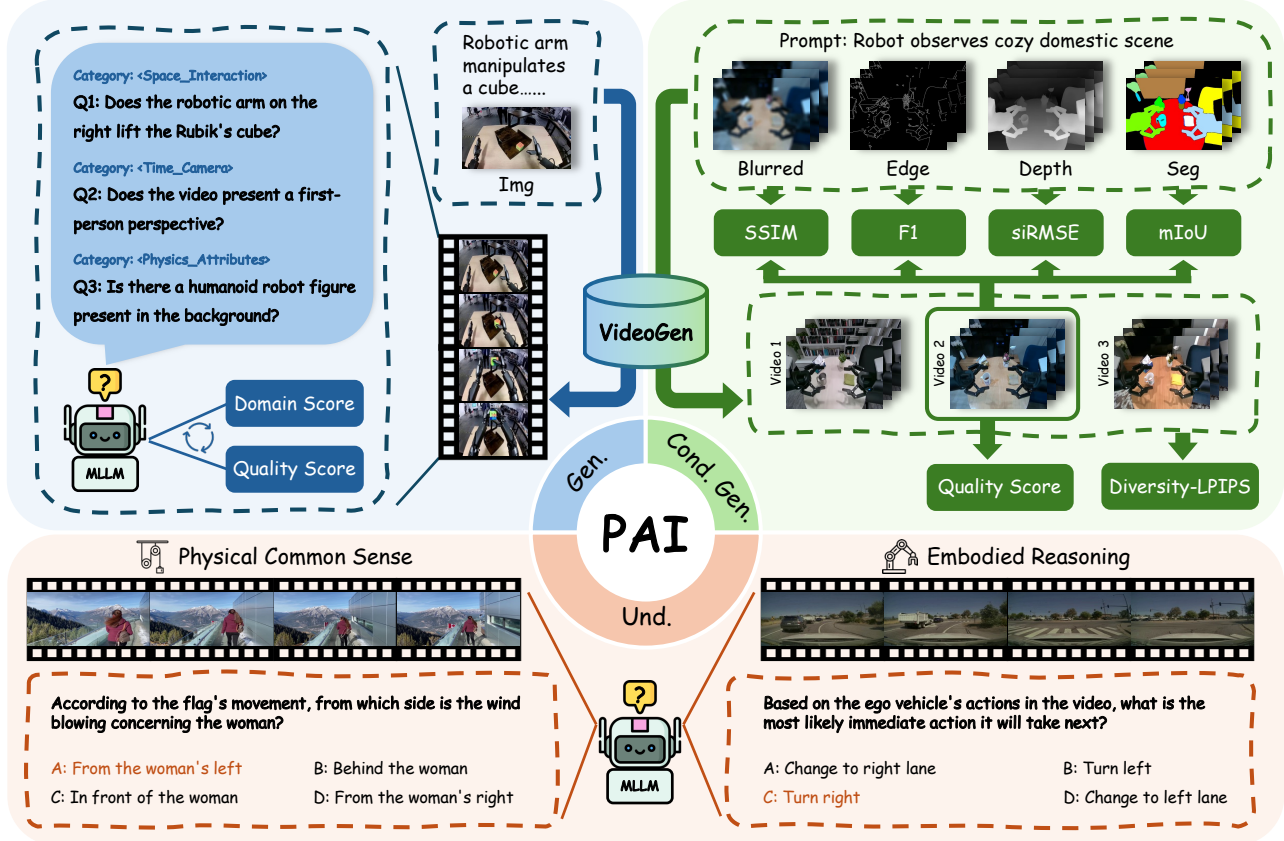


Figure 1. **PAI-Bench overview.** PAI-Bench is a real-world benchmark for evaluating Physical AI across video generation, conditional video generation, and video understanding, spanning robotics, autonomous driving, human actions, industry, and other physical domains.

Abstract

Physical AI aims to develop models that can perceive and predict real-world dynamics; yet, the extent to which current multi-modal large language models and video generative models support these abilities is insufficiently understood. We introduce Physical AI Bench (PAI-Bench), a unified and comprehensive benchmark that evaluates perception and prediction capabilities across video generation, conditional video generation, and video understanding, comprising 2,808 real-world cases with task-aligned metrics designed to capture physical plausibility and domain-

specific reasoning. Our study provides a systematic assessment of recent models and shows that video generative models, despite strong visual fidelity, often struggle to maintain physically coherent dynamics, while multi-modal large language models exhibit limited performance in forecasting and causal interpretation. These observations suggest that current systems are still at an early stage in handling the perceptual and predictive demands of Physical AI. In summary, PAI-Bench establishes a realistic foundation for evaluating Physical AI and highlights key gaps that future systems must address. The code is available at <https://github.com/SHI-Labs/physical-ai-bench>.

1. Introduction

“Human intelligence was not forged in a vacuum. It was sculpted by the relentless demands of the physical world.”

We learned to think because we had to move, grasp, and survive. AI systems stand at a similar threshold. To mature from a mere information processor into a truly capable partner, they must leave the clean, predictable confines of digital simulation and enter the messy, unforgiving reality. This is the core mission of **Physical AI**: the embodiment of intelligent algorithms into autonomous systems that must *perceive, predict, and act* directly within the physical world.

In this work, we focus on the *perception* and *prediction* components of this mission through visual signals, specifically video streams. Regarding *perception*, while recent multi-modal large language models (MLLMs) have demonstrated escalating capabilities with increased scale [8, 50, 74], their advancements are predominantly validated using benchmarks for abstract reasoning (e.g., OCR [30], mathematical problems [71]) and simple daily-life perception [28, 69, 100]. Their true efficacy within specialized Physical AI applications remains largely uncharted. Similarly, in the domain of *prediction*, a Physical AI agent must rely on an internal *world model* to forecast the consequences of actions and complex physical dynamics. Video generative models (VGMs) [16, 62] represent a powerful new paradigm for learning such dynamics, as they are implicitly trained to understand physics rules by predicting future video frames. However, existing benchmarks [38, 39, 52, 99] for these models have largely focused on assessing aesthetic appeal and temporal consistency, with far less attention paid to evaluating their fundamental understanding of real-world rules. It thus remains unclear whether they can model complex interactions, adhere to physical laws, or make reasonable predictions in practical applications.

To address these deficiencies and advance Physical AI, we introduce Physical AI Bench (**PAI-Bench**), a comprehensive benchmark designed to systematically evaluate model performance on tasks grounded in Physical AI. PAI-Bench is structured into three distinct tracks:

- **PAI-Bench-G** (Generation): Targets *prediction* by evaluating visual quality and physical plausibility of VGMs.
- **PAI-Bench-C** (Conditional Generation): Further probes *prediction* by evaluating the fidelity of conditional VGMs, specifically their conformance to input control signals (e.g., depth maps).
- **PAI-Bench-U** (Understanding): Targets the *perception* component by evaluating MLLM capabilities on video understanding tasks specific to the Physical AI domain.

All three tracks adhere to a unified design principle: grounding evaluation in physically meaningful tasks and real-world data. Specifically, all videos are sourced from real-world captures, such as dashcam recordings. Further-

more, PAI-Bench covers several sub-domains, such as autonomous vehicles, robotics, ego-view scenes, *etc.*, across three tracks, comprising 2,808 high-quality cases.

Based on PAI-Bench, we conducted a comprehensive evaluation of leading VGMs and MLLMs. Extensive experiments reveal that while SOTA VGMs like Veo3 [21] achieve high visual fidelity, their *prediction* abilities are limited, as they often fail to adhere to basic physical laws or to model complex real-world dynamics. Concurrently, the *perception* capabilities of powerful MLLMs, such as GPT-5 [61], lag substantially behind human performance. These findings suggest that the core *prediction* and *perception* capabilities required for Physical AI remain in a nascent stage.

In summary, our contributions are three-fold:

- We introduce PAI-Bench, the first benchmark to provide a unified and comprehensive review of video generation, conditional video generation, and video understanding.
- As the first high-quality benchmark grounded in Physical AI, PAI-Bench spans diverse scenarios (e.g., industrial environments, everyday activities, physical common sense) and incorporates a suite of task-aligned metrics.
- We conduct a large-scale evaluation of 15 VGMs, 4 conditional VGMs with 5 control settings, and 16 MLLMs, establishing the current landscape of Physical AI and highlighting key challenges to guide future research.

2. Related Work

Prediction and *Perception* are foundational capabilities for Physical AI applications. However, a significant gap persists in current benchmarking: the absence of a comprehensive framework specifically for the Physical AI domain. Furthermore, existing benchmarks are largely fragmented, typically assessing prediction or perception in isolation.

In the realm of *prediction*, world models [4, 12, 16, 34] are increasingly vital for applications such as robotics and autonomous driving. A prominent paradigm utilizes VGMs to forecast high-fidelity future frames [1, 4, 17, 21, 36, 60, 62, 77, 86, 91, 95]. This progress has spurred the proposal of numerous benchmarks [26, 41, 52, 53, 56, 96]. However, existing evaluation methodologies are often narrow in scope. Some benchmarks [38, 39, 99] primarily assess visual quality and temporal consistency, while others [11, 14, 22, 56] concentrate exclusively on physical plausibility. We introduce **PAI-Bench-G**, which distinguishes itself by integrating both aspects via a domain score and a quality score, with a strong emphasis on Physical AI applications. Furthermore, a significant evaluation gap has emerged as VGMs [3, 4, 21, 36, 62, 77, 95] increasingly leverage multimodal control signals (e.g., depth maps) for guided synthesis. No existing benchmark systematically tests this controllability. **PAI-Bench-C** is therefore introduced as the first benchmark to systematically evaluate this condition guided

Table 1. **Comparison with Other Benchmarks.** Our PAI-Bench is designed with comprehensive domains in Physical AI. Specifically, we focus on scenarios involving practical applications, such as autonomous vehicles (AV), industry, embodied AI, and ego-centric views.

Benchmarks	# Examples	Video Generation	Conditional Video Generation	Video Understanding	Embodied AI	AV	Industry	Egocentric Views	Physical Common Sense
EvalCrafter [52]	700	✓	✗	✗	✗	✗	✗	✗	✗
VBench [38]	800	✓	✗	✗	✗	✗	✗	✗	✗
Physics-IQ [58]	198	✓	✗	✗	✗	✗	✗	✗	✓
EgoSchema [55]	5000	✗	✗	✓	✗	✗	✗	✓	✗
VideoMME [29]	2700	✗	✗	✓	✗	✗	✗	✓	✗
MVP Bench [44]	54800	✗	✗	✓	✓	✗	✗	✓	✓
CausalVQA [27]	793	✗	✗	✓	✗	✗	✗	✓	✗
PhyGenBench [56]	160	✓	✗	✗	✗	✗	✗	✗	✓
IntPhys2 [14]	1070	✓	✗	✓	✓	✗	✗	✗	✓
PAI-Bench (Ours)	2808	✓	✓	✓	✓	✓	✓	✓	✓

generation capability. Similarly, in the *perception* domain, while numerous benchmarks exist for general video understanding (e.g., action recognition) [28, 37, 48, 69, 71, 89, 100], evaluations for physically grounded reasoning remain notably limited, particularly for video-based analysis [20, 27, 47, 58, 72]. **PAI-Bench-U** bridges this gap by unifying tasks for general video comprehension with a dedicated suite of physically grounded challenges.

In summary, PAI-Bench is the first comprehensive benchmark to focus specifically on the Physical AI domain. By evaluating the capabilities of *prediction* and *perception*, it provides a holistic framework for assessing the preparedness of AI models for real-world physical interaction.

3. PAI-Bench

Overview. This section details the foundational principles and construction process for each track of PAI-Bench. We elaborate on the core motivation, the design principles, and the data curation process. Specifically, PAI-Bench-G is presented in Section 3.1, PAI-Bench-C in Section 3.2, and PAI-Bench-U in Section 3.3.

3.1. PAI-Bench-G: Video Generation

For *prediction* capability, models should be able to forecast the next physically plausible states based on existing information to dynamically adjust their policies. Within the current community, VGMs are considered the most promising approach for this task. To evaluate the efficacy of these models, we propose two metrics: 1) *Quality Score* to assess the fidelity and coherence of the generated video, and 2) *Domain Score* to validate physical plausibility.

Data Curation. We curated videos from open-source datasets and public web sources, focusing on Physical AI domains (e.g., autonomous driving, robotics, and industrial applications). The annotation involved a two-stage process. First, to generate high-fidelity prompts, we employed an advanced MLLM, i.e., Qwen2.5-VL-72B-Instruct [10], for initial video captioning, followed by rigorous manual curation and correction. Second, to support *Domain Score* assessment, we generated QA pairs by leveraging the on-

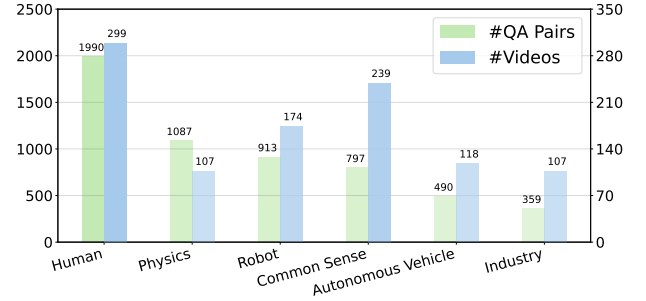


Figure 2. **Numbers of videos and QA pairs across domains of PAI-Bench-G.** QA pairs are used for *Domain Score* evaluation.

tology from [7]. This process used an MLLM for candidate generation, with subsequent manual refinement ensuring accuracy and relevance. The final dataset comprises 1,044 video-prompt pairs and 5,636 QA pairs across 6 domains, as illustrated in Figure 2 and Figure 3.

Quality Score. We adopt 8 metrics from VBench and VBench++ [38, 39, 99]. Among them, subject consistency, background consistency, motion smoothness, aesthetic quality, imaging quality, and overall consistency evaluate the quality of the generated video and its alignment with the textual input. In addition, image-to-video subject and image-to-video background assess the quality of image-conditioned video generation.

Domain Score. Evaluating the physical plausibility of the generated video necessitates analyzing fine-grained spatio-temporal details against foundational world knowledge. To address this evaluation, we employ Qwen3-VL-235B-A22B-Instruct [9] as a judge. We utilize the curated QA pairs, which define the expected physical and semantic content, to query the MLLM against the generated video. The Domain Score is defined as the MLLM’s response accuracy across this QA set, quantifying the video’s adherence to the specified physical and semantic constraints.

3.2. PAI-Bench-C: Conditional Video Generation

Control signals provide an interface through which humans can constrain the solution trajectory of VGMs, making them highly valuable in practical applications. In conditional

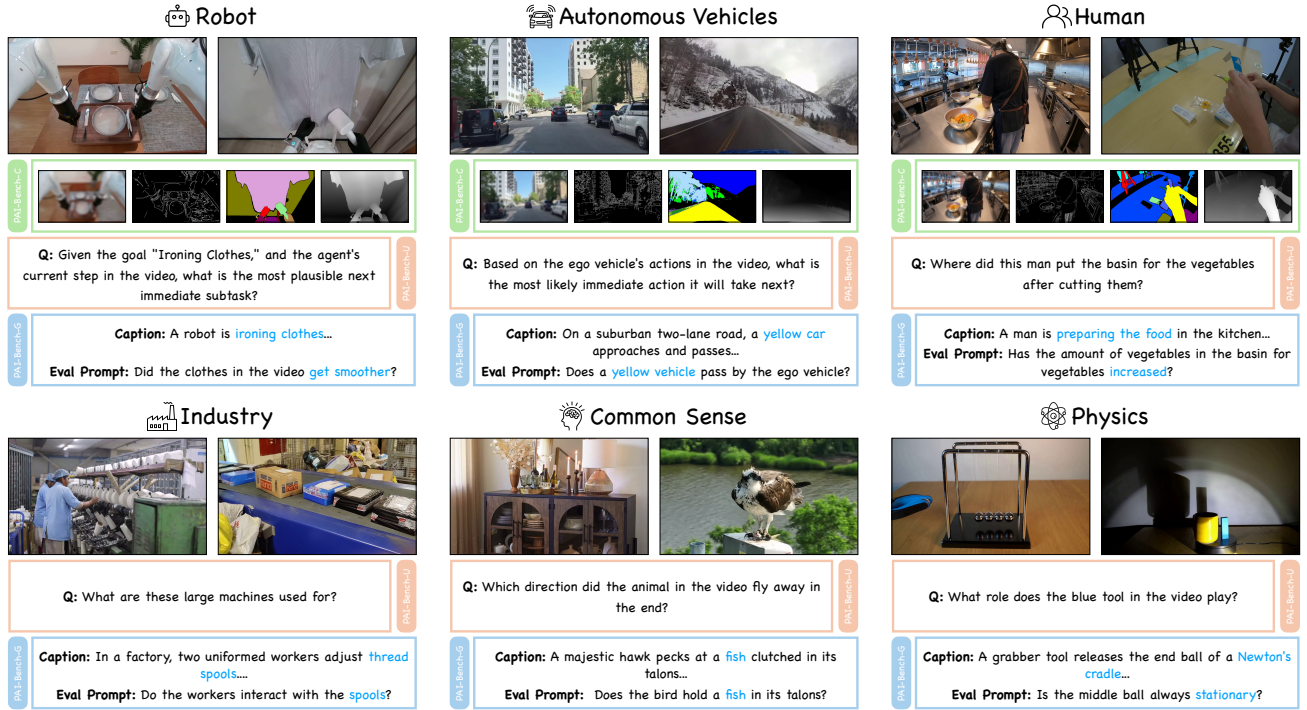


Figure 3. **Examples of PAI-Bench.** PAI-Bench focuses on Physical AI application scenarios across six domains, where only the first frame of each video is shown for brevity. For **PAI-Bench-C**, we present the blurred, edge, segmentation, and depth videos that serve as control signals. For **PAI-Bench-U**, we show the questions used for video understanding. For **PAI-Bench-G**, we show the input captions used for generation and a derived prompt used for *Domain Score* evaluation.

generation, an ideal outcome should satisfy three key criteria: 1) faithfulness to the control signals, 2) visual quality of the generated videos, and 3) diversity in the generated results under the same configurations of control signals.

Evaluation Metrics. Based on these principles, we evaluate faithfulness using a suite of fidelity metrics: *Blur SSIM*, *Edge F1*, *Depth si-RMSE*, and *Mask mIoU*. These metrics respectively quantify adherence to blur, edge, depth, and segmentation map controls. The computation projects the generated video into the specific modality space, utilizing Blur Kernel [83], Canny [75], Video Depth Anything [18], and GroundingDINO [51] with SAM2 [65], and then measures the similarity to the ground truth control signal. Visual quality is measured using DOVER [87, 88, 90], and diversity is assessed with LPIPS [98]. Detailed calculation protocols are presented in the supplementary materials.

Dataset Curation. We curated 600 videos by sampling 200 clips from each of the three datasets: AgiBot [2, 70] for robotics, OpenDV [93] for autonomous driving, and Ego-Exo-4D [33] for ego-centric tasks. Ground-truth control signals were extracted using the modality-specific models described above. To enable diversity evaluation, similar to PAI-Bench-G, we employ an advanced MLLM, *i.e.*, Qwen2.5-VL-72B-Instruct [10], to generate high-fidelity prompts followed by manual refinement. Based on these captions, we further use the MLLM to modify the visu-

ally dominant objects in the initial caption, with additional manual refinement, so that the resulting prompts describe coherent yet novel scenes. For each video, this process produces one caption corresponding to the original video and five novel captions.

3.3. PAI-Bench-U: Physical Video Understanding

Human cognition leverages multimodal sensory integration to build robust representations of the physical world. This *perception* capability is also foundational for Physical AI. Focusing on the interpretation of visual signals from videos, we posit that robust Physical AI models must possess two critical capabilities: 1) *Physical Common Sense Reasoning* and embodiment-agnostic understanding of environmental physics to assess real world plausibility, and 2) *Embodied Reasoning*, the capacity for an agent to perceive, reason about, and plan future interactions.

Physical Common Sense Reasoning. Humans acquire physical common sense primarily through passive observation [66], forming an implicit knowledge base of real-world plausibility. To formalize this concept, we delineate an ontology comprising three domains:

- *Space* governs object relations, interactions, and environmental context. This includes determining spatial feasibility and understanding scene composition.
- *Time* concerns actions and events unfolding over a dura-

Table 3. **Evaluation results of 15 VGMs on PAI-Bench-G.** Metrics in Domain Score: Common Sense (CS), Autonomous Vehicle (AV), Robot (RO), Industry (IN), Human (HU), Physics (PH); and Quality Score: Subject Consistency (SC), Background Consistency (BC), Motion Smoothness (MS), Aesthetic Quality (AQ), Imaging Quality (IQ), Overall Consistency (OC), I2V Subject (IS), I2V Background (IB). Blue means the best across open-source models.

Models	Overall	Domain Score							Quality Score								
		CS	AV	RO	IN	HU	PH	Avg.	SC	BC	MS	AQ	IQ	OC	IS	IB	Avg.
Source Videos	83.9	96.6	78.7	91.8	89.0	86.6	93.1	89.8	93.3	94.2	99.1	51.7	68.4	21.5	97.8	98.2	78.0
Proprietary Models																	
Veo3 [21]	82.2	93.0	72.1	88.8	89.0	84.4	89.7	86.8	91.4	93.1	99.2	51.9	69.8	21.7	97.0	96.9	77.6
Open-source Models																	
DynamicCrafter [91]	68.3	73.9	43.9	48.9	70.9	60.7	80.8	63.0	91.1	92.5	94.9	51.5	68.0	21.2	84.5	86.2	73.7
LTX-Video-2B [35]	77.5	88.1	60.5	69.4	82.6	77.1	86.2	78.0	89.2	92.7	98.7	53.2	71.3	21.1	95.0	95.9	77.1
MAGI-1-4.5B [73]	77.5	88.3	63.4	71.3	82.5	77.3	86.3	78.7	92.1	93.3	99.0	50.4	61.8	21.6	94.5	98.1	76.3
HunyuanVideo-I2V [43]	77.6	86.7	60.3	68.9	82.4	75.8	85.8	77.1	94.3	95.3	99.5	52.1	65.2	21.5	98.6	97.6	78.0
CogVideoX1.5 [95]	78.4	88.7	64.4	73.6	84.6	79.1	88.7	80.3	91.6	93.9	98.5	50.0	66.5	21.2	95.0	96.1	76.6
LTX-Video-13B [35]	78.4	88.6	60.0	71.8	84.6	79.1	88.8	79.5	90.6	93.5	99.0	53.5	69.5	21.4	95.7	96.0	77.4
CogVideoX [95]	78.6	88.4	62.7	75.3	85.0	81.0	88.5	80.9	91.4	93.4	98.0	51.2	64.6	21.3	94.1	95.9	76.3
MAGI-1-24B [73]	79.4	91.4	68.0	76.5	85.4	81.5	87.0	82.4	90.0	92.4	99.0	50.2	64.2	21.4	96.8	97.9	76.5
Cosmos-Predict2-2B [4]	79.7	91.2	70.7	81.9	86.4	83.1	88.1	84.2	88.7	92.1	97.6	49.3	65.9	21.6	91.9	94.6	75.2
Cosmos-Predict2-14B [4]	80.0	92.5	69.8	81.3	86.6	82.4	89.2	84.3	89.6	92.8	98.0	49.8	67.5	21.5	92.2	94.9	75.8
Wan2.1-I2V-14B [77]	80.8	92.5	69.4	79.6	87.2	83.4	89.5	84.3	90.4	93.0	98.0	51.7	70.0	21.4	96.5	97.2	77.3
Wan2.2-TI2V-5B [77]	81.3	92.2	69.1	82.9	86.4	83.3	90.3	84.7	92.0	93.9	98.8	52.3	69.6	21.4	97.4	98.0	77.9
Cosmos-Predict2.5-2B [4]	81.4	91.5	70.4	82.7	86.8	83.5	91.7	84.9	92.3	94.2	99.1	52.8	70.7	21.2	96.2	97.2	78.0
Wan2.2-I2V-A14B [77]	82.3	94.1	73.0	86.8	88.6	84.5	92.7	87.1	91.2	93.2	98.3	51.8	69.2	21.5	97.2	97.8	77.5

Table 4. **Evaluation results of 4 conditional VGMs on PAI-Bench-C.** For each model, the control signal settings consist of either a single video or a combination of multiple signal videos. Green means the best across control signal settings for each model.

Model	Condition	Blur SSIM $\uparrow$	Edge F1 $\uparrow$	Depth si-RMSE $\downarrow$	Mask mIoU $\uparrow$	Quality Score $\uparrow$	Diversity $\uparrow$
Cosmos-Transfer1-7B [3]	Blur	0.89	0.20	0.66	0.73	6.56	0.19
	Edge	0.77	0.38	0.85	0.73	6.76	0.28
	Depth	0.67	0.15	0.76	0.71	6.89	0.39
	Seg	0.62	0.11	1.13	0.70	6.02	0.42
	All	0.82	0.26	0.70	0.74	9.24	0.22
Cosmos-Transfer2.5-2B [4]	Blur	0.91	0.26	0.54	0.75	8.77	0.18
	Edge	0.76	0.39	0.74	0.74	8.05	0.36
	Depth	0.70	0.17	0.83	0.72	7.30	0.41
	Seg	0.66	0.13	1.07	0.71	7.87	0.44
	All	0.90	0.45	0.59	0.77	9.24	0.13
Wan2.2-Fun-5B-Control [5]	Edge	0.61	0.27	1.01	0.71	8.79	0.40
	Depth	0.56	0.11	1.82	0.62	9.32	0.48
Wan2.2-Fun-A14B-Control [5]	Blur	0.57	0.09	2.11	0.50	8.81	0.53
	Edge	0.68	0.37	0.84	0.74	9.00	0.38
	Depth	0.56	0.11	2.10	0.58	9.22	0.52
	Seg	0.47	0.10	1.60	0.66	7.79	0.36

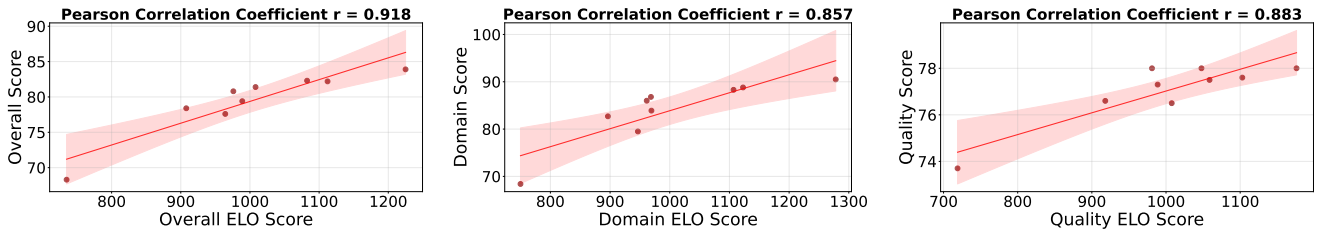


Figure 6. **Pearson correlation analysis** on PAI-Bench-G. The red shaded regions indicate the 0.95 confidence intervals.

4.2. PAI-Bench-C

Models. Our evaluation targets two families of controllable video generation models: Cosmos-Transfer [3, 4] and Wan-Fun [5].

Evaluation Suites. We standardize generation settings across all models, configuring them to produce 121-frame

videos, preserve the input video’s aspect ratio, and use their default configurations. We assess model performance using five distinct conditioning signals: blurred video (*Blur*), edge video (*Edge*), depth video (*Depth*), segmentation maps (*Seg.*), and a combination of all signals with equal weighting (*All*). The *All* condition is exclusively evaluated on

Table 5. **Evaluation of 16 MLLMs on PAI-Bench-U**. Embodied reasoning domains: BridgeData (BD), RoboVQA (RV), RoboFail (RF), Agibot (AB), HoloAssist (HA), Autonomous Vehicle (AV). Red denotes the best result across either proprietary or open-source models.

Models	Overall	Common Sense				Embodied Reasoning						
		Space	Time	Physics	Avg.	BD	RV	RF	AB	HA	AV	Avg.
Human Performance	93.2	92.1	94.3	95.0	93.6	94.0	95.5	97.0	90.0	89.0	100.0	88.9
Random Guess	37.0	41.5	38.8	38.3	38.9	25.0	50.0	50.0	26.9	25.0	32.5	35.2
<i>Proprietary Models</i>												
Claude-3.5-Sonnet [6]	46.0	55.0	46.6	46.9	47.8	29.0	74.5	58.0	28.0	38.0	34.0	44.1
GPT-4o [40]	56.2	61.2	57.0	59.7	58.6	44.0	68.2	71.0	45.0	55.0	38.0	53.8
GPT-5 [61]	61.8	63.8	65.8	61.5	63.9	38.0	85.5	72.0	45.0	72.0	43.0	59.7
<i>Open-source Models</i>												
InternVL3.5-14B [80]	48.8	50.0	51.3	47.3	49.7	23.0	80.0	67.0	27.0	56.0	31.0	47.9
InternVL3.5-30B-A3B [80]	49.4	48.8	55.7	46.0	51.2	37.0	74.5	60.0	23.0	55.0	34.0	47.7
InternVL3.5-8B [80]	49.7	50.0	55.0	44.7	50.5	29.0	75.5	63.0	39.0	49.0	35.0	48.9
Qwen2.5-VL-7B [10]	51.0	51.2	48.7	39.8	45.7	35.0	87.3	63.0	53.0	60.0	36.0	56.2
Qwen2.5-VL-32B [10]	55.3	50.0	62.1	48.7	55.5	35.0	93.6	65.0	45.0	56.0	32.0	55.1
Cosmos-Reason1-7B [7]	55.7	63.8	55.7	46.0	53.1	41.0	91.8	66.0	41.0	59.0	47.0	58.2
InternVL3.5-38B [80]	55.8	58.8	59.7	49.6	55.8	36.0	82.7	66.0	44.0	69.0	34.0	55.7
InternVL3.5-241B-A28B [80]	56.3	60.0	57.4	53.5	56.3	34.0	78.2	66.0	43.0	75.0	40.0	56.4
Qwen3-VL-8B [9]	56.8	53.8	58.4	50.9	55.0	35.0	89.1	61.0	49.0	75.0	40.0	58.7
Qwen3-VL-30B-A3B [9]	59.5	52.5	60.4	58.4	58.6	38.0	90.9	69.0	41.0	73.0	47.0	60.3
Qwen2.5-VL-72B [10]	60.8	65.0	57.7	57.5	58.6	50.0	91.8	68.0	52.0	70.0	43.0	63.0
Qwen3-VL-32B [9]	62.0	53.8	67.8	59.7	62.9	42.0	90.9	71.0	50.0	72.0	38.0	61.1
Qwen3-VL-235B-A22B [9]	64.7	56.2	69.5	61.9	64.9	42.0	93.6	71.0	45.0	76.0	56.0	64.4

Cosmos-Transfer [3]. For each control signal, we generate six videos using six unique captions, maintaining a fixed random seed. Notably, since Wan2.2-Fun-5B-Control [5] failed to produce coherent results when conditioned on blur or segmentation maps, these are omitted from our evaluation. All results are summarized in Table 4.

Multi-signal conditioning enhances video quality. The results in Table 4 clearly demonstrate the benefits of multi-signal conditioning. For both Cosmos-Transfer [3] models, the *All* condition achieves the highest Quality Score, significantly outperforming any single control signal. This finding suggests a valuable practical application: instead of relying on a single degraded input, such as a blurry or noisy video, users can first extract complementary control signals from the source. By providing the model with this comprehensive set of control signals, it becomes possible to reconstruct a video of substantially higher quality.

Segmentation signals induce poor segmentation fidelity. For Cosmos-Transfer [3] models, surprisingly, we observe that using segmentation maps as the control signal results in the lowest Mask mIoU. We hypothesize that this stems from segmentation maps being the noisiest supervision signal among all conditions. Even state-of-the-art segmentation models, *e.g.*, SAM2 [65], can produce masks with lower temporal consistency than other control signals, such as the occasional missing object masks across frames. Consequently, the resulting training supervision is likely to be highly unreliable, degrading the model’s ability.

4.3. PAI-Bench-U

Models. We evaluate 16 MLLMs, covering both proprietary and open-source models. The proprietary models include Claude-3.5-Sonnet [6], GPT-4o [40], and GPT-5 [61]. For GPT-5, we use minimal reasoning effort. The open-source group spans several major families, including Qwen2.5-VL [10], Qwen3-VL [9], Cosmos-Reason1 [7], and InternVL3.5 [80]. This collection covers a broad spectrum of model scales, from 7B to over 200B. To analyze model performance, we also establish a human baseline derived from an annotator team.

Evaluation Suites. We conduct our evaluation using the LMMs-Eval framework [97]. All models are evaluated using 16 video frames as input, except for InternVL3.5, which can accommodate only 8 frames due to its context length. We follow the method in [79] to adaptively resize video frames. For all model outputs, we apply the lightweight LLM Qwen3-8B [92] as a post-processing model to extract the final letter option from the full responses. Reasoning is off by default. All results are summarized in Table 5.

PAI-Bench-U is unbiased. Prevailing video understanding benchmarks often suffer from significant biases [32, 46, 64, 74]. Specifically, models frequently 1) exploit strong language priors to infer answers without grounding in visual data, and 2) leverage a *static bias*, where questions are answerable from a single frame while obviating the need for temporal context. We analyzed PAI-Bench-U for these issues by evaluating SOTA models with varying input frame counts in Figure 7. First, with zero frames (*i.e.*,

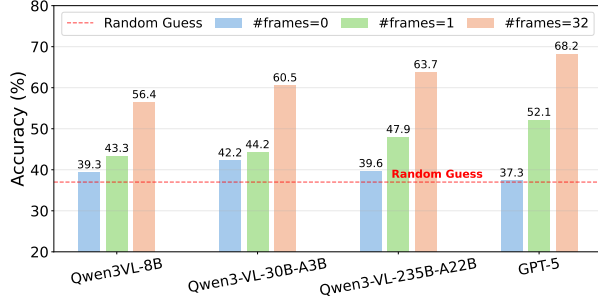


Figure 7. Comparison across different frame counts.

text-only input), model performance degraded to the level of random guessing, which confirms that PAI-Bench-U effectively neutralizes language priors, ensuring that performance gains are sourced from visual understanding. Furthermore, we observed a substantial performance gap between the 1-frame and 32-frame inputs. This indicates that the tasks are not solvable with static information alone, validating PAI-Bench-U’s requirement for temporal context comprehension.

MLLMs lag significantly behind human performance, with no clear proprietary model advantage. The results in Table 5 highlight a substantial performance gap between all models and human accuracy (93.2). This high performance validates PAI-Bench-U as a well-posed benchmark; yet, the stark discrepancy underscores the significant limitations of current models in understanding capabilities in the Physical AI domain. Surprisingly, proprietary models do not uniformly outperform open-source models; for instance, Qwen3-VL-235B [9] achieves a higher accuracy than GPT-5 [61]. This finding suggests that the Physical AI domain has not yet been a primary data collection or optimization priority for the wider field.

Textual thinking does not guarantee performance gains. We investigated whether introducing an additional thinking process during inference enhances performance. As presented in Table 6, the Qwen3-VL [9] series, which employs textual-only thinking, unexpectedly exhibited a slight performance degradation, particularly in embodied reasoning tasks. We attribute this to the high demands of these tasks, which require perceiving fine-grained visual details, such as minute regions or subtle temporal changes. When vision modules fail to capture these nuances, subsequent textual reasoning operates without the necessary grounding, rendering it ineffective. This hypothesis is supported by GPT-5 [61]’s results. We use medium reasoning effort for GPT-5, and the model shows clear performance gains under this setting. Crucially, GPT-5’s reasoning involves both textual and visual thinking. This integrated visual reasoning appears to be an effective solution for capturing fine-grained information. Therefore, this finding underscores the necessity of advancing visual thinking capabilities within the Physical AI domain.

Table 6. Ablation on think mode of MLLMs on PAI-Bench-U.

Models	Thinking Mode	Overall	Common Sense	Embodied Reasoning
Qwen3-VL-8B [9]	✗	56.8	55.0	58.7
	✓	57.3 (+0.5)	57.0 (+2.0)	57.7 (-1.0)
Qwen3-VL-30B-A3B [9]	✗	59.5	58.6	60.3
	✓	57.3 (-2.2)	56.1 (-2.5)	58.5 (-1.8)
Qwen3-VL-32B [9]	✗	62.0	62.9	61.1
	✓	61.0 (-1.0)	63.7 (+0.8)	58.4 (-2.7)
Qwen3-VL-235B-A22B [9]	✗	64.7	64.9	64.4
	✓	63.7 (-1.0)	66.4 (+1.5)	61.0 (-3.4)
GPT-5 [61]	✗	61.8	63.9	59.7
	✓	69.8 (+8.0)	71.4 (+7.5)	68.2 (+8.5)

5. Further Analysis and Discuss

Based on an in-depth analysis and cross-comparison of three splits of PAI-Bench, we extract several common insights and subsequently synthesize a central conclusion:

Physical AI remains a largely unsolved domain. In PAI-Bench-G, leading open-source VGMs, including Wan2.2-I2V-A14B [77] (81.4), achieve performance comparable to the proprietary Veo3 [21] (82.2), as shown in Table 3. A similar trend is observed for MLLMs, where the open-source Qwen3-VL-235B-A22B [9] (64.7) surpasses the proprietary GPT-5 [61] (61.8) by 2.9 points in Table 5. This limited differentiation, coupled with generally lower absolute performance on more challenging settings, suggests one of two systemic challenges. 1) A community-wide data gap may exist, as Physical AI has not been a primary target for data collection or optimization. This would necessitate a focused, large-scale data collection effort to advance the domain. 2) Alternatively, the data may exist, but both proprietary and open-source models are failing to learn the specific capabilities required by Physical AI. For instance, VGMs may not be learning the underlying principles of the physical world [19, 58, 86], while MLLMs may not be learning to leverage common sense to guide agent policies. Either scenario strongly indicates that the field of Physical AI remains in a nascent stage.

6. Conclusion

In this work, we take a step toward bridging the gap between perception, prediction, and physical reasoning by introducing PAI-Bench, a comprehensive benchmark for evaluating models in the context of the physical world for video generation, conditional video generation, and video understanding. Through its three complementary components, PAI-Bench systematically examines how current models perceive, predict, and reason about physical phenomena. Our large-scale evaluations reveal that, despite impressive visual quality, state-of-the-art models still struggle to capture basic physical regularities and causal interactions. We hope that PAI-Bench will serve as a foundational benchmark for advancing the next generation of Physical AI systems that not only see and generate but also truly understand and act within the real world.

Acknowledgements

We would like to thank NVIDIA Research, especially the Cosmos team, for their support, which led to the creation of PAI-Bench. We also thank Yin Cui, Jinwei Gu, Heng Wang, Prithvijit Chattopadhyay, Andrew Z. Wang, Imad El Hanafi, and Ming-Yu Liu for their valuable feedback and collaboration that helped shape the project. This research was supported in part by the National Science Foundation under Award #2427478 - CAREER Program and by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, under Award #2229873 - National AI Institute for Exceptional Education. This project was also partially supported by cyberinfrastructure resources and services provided by the Georgia Institute of Technology.

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, et al. “Cosmos World Foundation Model Platform for Physical AI”. In: *CoRR* abs/2501.03575 (2025). DOI: 10 . 48550 / ARXIV . 2501 . 03575. arXiv: 2501 . 03575. URL: <https://doi.org/10.48550/arXiv.2501.03575>.
- [2] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, et al. “AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems”. In: *CoRR* abs/2503.06669 (2025). DOI: 10 . 48550 / ARXIV . 2503 . 06669. arXiv: 2503 . 06669. URL: <https://doi.org/10.48550/arXiv.2503.06669>.
- [3] Hassan Abu Alhaija, José M. Álvarez, Maciej Bala, et al. “Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control”. In: *CoRR* abs/2503.14492 (2025). DOI: 10 . 48550 / ARXIV . 2503 . 14492. arXiv: 2503 . 14492. URL: <https://doi.org/10.48550/arXiv.2503.14492>.
- [4] Arslan Ali, Junjie Bai, Maciej Bala, et al. “World Simulation with Video Foundation Models for Physical AI”. In: *arXiv preprint arXiv:2511.00062* (2025).
- [5] Alibaba PAI. *Wan2.2-Fun: a Alibaba-PAI Collection*. <https://huggingface.co/collections/alibaba-pai/wan22-fun>. Accessed: 2025-11-12. 2025.
- [6] Anthropic. *Claude 3.5 Sonnet*. 2024. URL: <https://www.anthropic.com/news/claude-3-5-sonnet>.
- [7] Alisson G. Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, et al. “Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning”. In: *CoRR* abs/2503.15558 (2025). DOI: 10 . 48550 / ARXIV . 2503 . 15558. arXiv: 2503 . 15558. URL: <https://doi.org/10.48550/arXiv.2503.15558>.
- [8] Jinze Bai, Shuai Bai, Shusheng Yang, et al. “Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities”. In: *CoRR* abs/2308.12966 (2023). DOI: 10 . 48550 / ARXIV . 2308 . 12966. arXiv: 2308 . 12966. URL: <https://doi.org/10.48550/arXiv.2308.12966>.
- [9] Jinze Bai, Shuai Bai, Shusheng Yang, et al. “Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond”. In: *arXiv preprint arXiv:2308.12966* (2023). URL: <https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list>.
- [10] Shuai Bai, Keqin Chen, Xuejing Liu, et al. “Qwen2.5-VL Technical Report”. In: *CoRR* abs/2502.13923 (2025). DOI: 10 . 48550 / ARXIV . 2502 . 13923. arXiv: 2502 . 13923. URL: <https://doi.org/10.48550/arXiv.2502.13923>.
- [11] Hritik Bansal, Clark Peng, Yonatan Bitton, et al. “VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation”. In: *CoRR* abs/2503.06800 (2025). DOI: 10 . 48550 / ARXIV . 2503 . 06800. arXiv: 2503 . 06800. URL: <https://doi.org/10.48550/arXiv.2503.06800>.
- [12] Amir Bar, Gaoyue Zhou, Danny Tran, et al. “Navigation World Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 15791–15801. DOI: 10 . 1109 / CVPR52734 . 2025 . 01472. URL: https://openaccess.thecvf.com/content/CVPR2025/html/Bar%5C_Navigation%5C_World%5C_Models%5C_CVPR%5C_2025%5C_paper.html.
- [13] Suneel Belkhale, Tianli Ding, Ted Xiao, et al. *RT-H: Action Hierarchies Using Language*. 2024. arXiv: 2403 . 01823 [cs.RO]. URL: <https://arxiv.org/abs/2403.01823>.
- [14] Florian Bordes, Quentin Garrido, Justine T. Kao, et al. “IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments”. In: *CoRR* abs/2506.09849 (2025). DOI: 10 . 48550 / ARXIV . 2506 . 09849. arXiv: 2506 . 09849. URL: <https://doi.org/10.48550/arXiv.2506.09849>.
- [15] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: 2021.
- [16] Jun Cen, Chaohui Yu, Hangjie Yuan, et al. “WorldVLA: Towards Autoregressive Action World Model”. In: *CoRR* abs/2506.21539 (2025). DOI: 10 . 48550 / ARXIV . 2506 . 21539. arXiv: 2506 . 21539. URL: <https://doi.org/10.48550/arXiv.2506.21539>.

- [17] Haoxin Chen, Yong Zhang, Xiaodong Cun, et al. “VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 7310–7320. DOI: [10.1109/CVPR52733.2024.00698](https://doi.org/10.1109/CVPR52733.2024.00698). URL: <https://doi.org/10.1109/CVPR52733.2024.00698>.
- [18] Sili Chen, Hengkai Guo, Shengnan Zhu, et al. “Video Depth Anything: Consistent Depth Estimation for Super-Long Videos”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 22831–22840. DOI: [10.1109/CVPR52734.2025.02126](https://openaccess.thecvf.com/content/CVPR2025/html/Chen%5C_Video%5C_Depth%5C_Anything%5C_Consistent%5C_Depth%5C_Estimation%5C_for%5C_Super-Long%5C_Videos%5C_CVPR%5C_2025%5C_paper.html). URL: https://openaccess.thecvf.com/content/CVPR2025/html/Chen%5C_Video%5C_Depth%5C_Anything%5C_Consistent%5C_Depth%5C_Estimation%5C_for%5C_Super-Long%5C_Videos%5C_CVPR%5C_2025%5C_paper.html.
- [19] Zhe Chen, Weiyun Wang, Hao Tian, et al. “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites”. In: *Science China Information Sciences* 67.12 (2024), p. 220101.
- [20] Wei Chow, Jiageng Mao, Boyi Li, et al. “PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding”. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL: <https://openreview.net/forum?id=Q6a9W6kzv5>.
- [21] Google DeepMind. *Veo: a text-to-video generation system*. Tech. rep. Technical Report. Google DeepMind, 2025.
- [22] Haoyi Duan, Hong-Xing Yu, Sirui Chen, et al. “WorldScore: A Unified Evaluation Benchmark for World Generation”. In: *CoRR* abs/2504.00983 (2025). DOI: [10.48550/ARXIV.2504.00983](https://doi.org/10.48550/ARXIV.2504.00983). arXiv: [2504.00983](https://arxiv.org/abs/2504.00983). URL: <https://doi.org/10.48550/ARXIV.2504.00983>.
- [23] David Eigen, Christian Puhrsch, and Rob Fergus. “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network”. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, et al. 2014, pp. 2366–2374. URL: <https://proceedings.neurips.cc/paper/2014/hash/7bccfde7714alebdf06c5f4cea752c1-Abstract.html>.
- [24] Arpad E Elo. “The proposed uscf rating system, its development, theory, and applications”. In: *Chess Life* 22.8 (1967), pp. 242–247.
- [25] Yuming Fang, Hanwei Zhu, Yan Zeng, et al. “Perceptual Quality Assessment of Smartphone Photography”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 3674–3683. DOI: [10.1109/CVPR42600.2020.00373](https://openaccess.thecvf.com/content%5C_CVPR%5C_2020/html/Fang%5C_Perceptual%5C_Quality%5C_Assessment%5C_of%5C_Smartphone%5C_Photography%5C_CVPR%5C_2020%5C_paper.html). URL: https://openaccess.thecvf.com/content%5C_CVPR%5C_2020/html/Fang%5C_Perceptual%5C_Quality%5C_Assessment%5C_of%5C_Smartphone%5C_Photography%5C_CVPR%5C_2020%5C_paper.html.
- [26] Weixi Feng, Jiachen Li, Michael Saxon, et al. “TC-Bench: Benchmarking Temporal Compositionality in Text-to-Video and Image-to-Video Generation”. In: *CoRR* abs/2406.08656 (2024). DOI: [10.48550/ARXIV.2406.08656](https://doi.org/10.48550/ARXIV.2406.08656). arXiv: [2406.08656](https://arxiv.org/abs/2406.08656). URL: <https://doi.org/10.48550/ARXIV.2406.08656>.
- [27] Aaron Foss, Chloe Evans, Sasha Mitts, et al. “CausalVQA: A Physically Grounded Causal Reasoning Benchmark for Video Models”. In: *CoRR* abs/2506.09943 (2025). DOI: [10.48550/ARXIV.2506.09943](https://doi.org/10.48550/ARXIV.2506.09943). arXiv: [2506.09943](https://arxiv.org/abs/2506.09943). URL: <https://doi.org/10.48550/ARXIV.2506.09943>.
- [28] Chaoyou Fu, Yuhao Dai, Yongdong Luo, et al. “Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 24108–24118. DOI: [10.1109/CVPR52734.2025.02245](https://openaccess.thecvf.com/content/CVPR2025/html/Fu%5C_Video-MME%5C_The%5C_First-Ever%5C_Comprehensive%5C_Evaluation%5C_Benchmark%5C_of%5C_Multi-modal%5C_LLMs%5C_in%5C_CVPR%5C_2025%5C_paper.html). URL: https://openaccess.thecvf.com/content/CVPR2025/html/Fu%5C_Video-MME%5C_The%5C_First-Ever%5C_Comprehensive%5C_Evaluation%5C_Benchmark%5C_of%5C_Multi-modal%5C_LLMs%5C_in%5C_CVPR%5C_2025%5C_paper.html.
- [29] Chaoyou Fu, Yuhao Dai, Yongdong Luo, et al. “Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 24108–24118. DOI: [10.1109/CVPR52734.2025.02245](https://openaccess.thecvf.com/content/CVPR2025/html/Fu%5C_Video-MME%5C_The%5C_First-Ever%5C_Comprehensive%5C_Evaluation%5C_Benchmark%5C_of%5C_Multi-modal%5C_LLMs%5C_in%5C_CVPR%5C_2025%5C_paper.html). URL: https://openaccess.thecvf.com/content/CVPR2025/html/Fu%5C_Video-MME%5C_The%5C_First-Ever%5C_Comprehensive%5C_Evaluation%5C_Benchmark%5C_of%5C_Multi-modal%5C_LLMs%5C_in%5C_CVPR%5C_2025%5C_paper.html.
- [30] Ling Fu, Biao Yang, Zhebin Kuang, et al. “OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning”. In: *CoRR* abs/2501.00321 (2025). DOI: [10.48550/ARXIV.2501.00321](https://arxiv.org/abs/2501.00321).

- 48550/ARXIV.2501.00321. arXiv: 2501.00321. URL: <https://doi.org/10.48550/arXiv.2501.00321>.
- [31] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, et al. “DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, et al. 2023. URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/9f09f316a3eaf59d9ced5ffae97e0f-Abstract-Conference.html.
 - [32] Yash Goyal, Tejas Khot, Douglas Summers-Stay, et al. “Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6325–6334. DOI: 10.1109/CVPR.2017.670. URL: <https://doi.org/10.1109/CVPR.2017.670>.
 - [33] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. “Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 19383–19400. DOI: 10.1109/CVPR52733.2024.01834. URL: <https://doi.org/10.1109/CVPR52733.2024.01834>.
 - [34] David Ha and Jürgen Schmidhuber. “World Models”. In: (2018). DOI: 10.5281/ZENODO.1207631. URL: <https://zenodo.org/record/1207631>.
 - [35] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, et al. “LTX-Video: Realtime Video Latent Diffusion”. In: *CoRR abs/2501.00103* (2025). DOI: 10.48550/ARXIV.2501.00103. arXiv: 2501.00103. URL: <https://doi.org/10.48550/arXiv.2501.00103>.
 - [36] Wenyi Hong, Ming Ding, Wendi Zheng, et al. “CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers”. In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: <https://openreview.net/forum?id=rB6TpjAuSRy>.
 - [37] Kairui Hu, Penghao Wu, Fanyi Pu, et al. “Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos”. In: *CoRR abs/2501.13826* (2025). DOI: 10.48550/ARXIV.2501.13826. arXiv: 2501.13826. URL: <https://doi.org/10.48550/arXiv.2501.13826>.
 - [38] Ziqi Huang, Yinan He, Jiashuo Yu, et al. “VBench: Comprehensive Benchmark Suite for Video Generative Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 21807–21818. DOI: 10.1109/CVPR52733.2024.02060. URL: <https://doi.org/10.1109/CVPR52733.2024.02060>.
 - [39] Ziqi Huang, Fan Zhang, Xiaojie Xu, et al. “VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models”. In: *CoRR abs/2411.13503* (2024). DOI: 10.48550/ARXIV.2411.13503. arXiv: 2411.13503. URL: <https://doi.org/10.48550/arXiv.2411.13503>.
 - [40] Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. “GPT-4o System Card”. In: *CoRR abs/2410.21276* (2024). DOI: 10.48550/ARXIV.2410.21276. arXiv: 2410.21276. URL: <https://doi.org/10.48550/arXiv.2410.21276>.
 - [41] Pengliang Ji, Chuyang Xiao, Huilin Tai, et al. “T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*. IEEE, 2024, pp. 5325–5335. DOI: 10.1109/CVPRW63382.2024.00541. URL: <https://doi.org/10.1109/CVPRW63382.2024.00541>.
 - [42] Junjie Ke, Qifei Wang, Yilin Wang, et al. “MUSIQ: Multi-scale Image Quality Transformer”. In: *CoRR abs/2108.05997* (2021). arXiv: 2108.05997. URL: <https://arxiv.org/abs/2108.05997>.
 - [43] Weijie Kong, Qi Tian, Zijian Zhang, et al. “Hunyuan-Video: A Systematic Framework For Large Video Generative Models”. In: *CoRR abs/2412.03603* (2024). DOI: 10.48550/ARXIV.2412.03603. arXiv: 2412.03603. URL: <https://doi.org/10.48550/arXiv.2412.03603>.
 - [44] Benno Krojer, Mojtaba Komeili, Candace Ross, et al. “A Shortcut-aware Video-QA Benchmark for Physical Understanding via Minimal Video Pairs”. In: *CoRR abs/2506.09987* (2025). DOI: 10.48550/ARXIV.2506.09987. arXiv: 2506.09987. URL: <https://doi.org/10.48550/arXiv.2506.09987>.
 - [45] LAION-AI. *aesthetic-predictor*. <https://github.com/LAION-AI/aesthetic-predictor>. 2022.
 - [46] Jie Lei, Tamara L. Berg, and Mohit Bansal. “Revealing Single Frame Bias for Video-and-Language Learning”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics, 2023, pp. 487–507. DOI: 10.18653/V1/2023.ACL-LONG.29. URL: <https://doi.org/10.18653/v1/2023.acl-long.29>.

- [47] Jialuo Li, Wenhao Chai, Xingyu Fu, et al. “Science-T2I: Addressing Scientific Illusions in Image Synthesis”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 2734–2744. DOI: [10.1109/CVPR52734.2025.00261](https://doi.org/10.1109/CVPR52734.2025.00261). URL: https://openaccess.thecvf.com/content/CVPR2025/html/Li%5C_Science-T2I%5C_Addressing%5C_Scientific%5C_Illusions%5C_in%5C_Image%5C_Synthesis%5C_CVPR%5C_2025%5C_paper.html.
- [48] Kunchang Li, Yali Wang, Yinan He, et al. “MVBench: A Comprehensive Multi-modal Video Understanding Benchmark”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 22195–22206. DOI: [10.1109/CVPR52733.2024.02095](https://doi.org/10.1109/CVPR52733.2024.02095). URL: <https://doi.org/10.1109/CVPR52733.2024.02095>.
- [49] Zhen Li, Zuo-Liang Zhu, Linghao Han, et al. “AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 9801–9810. DOI: [10.1109/CVPR52729.2023.00945](https://doi.org/10.1109/CVPR52729.2023.00945). URL: <https://doi.org/10.1109/CVPR52729.2023.00945>.
- [50] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. *Visual Instruction Tuning*. 2023.
- [51] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. “Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection”. In: *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*. Ed. by Ales Leonardis, Elisa Ricci, Stefan Roth, et al. Vol. 15105. Lecture Notes in Computer Science. Springer, 2024, pp. 38–55. DOI: [10.1007/978-3-031-72970-6_3](https://doi.org/10.1007/978-3-031-72970-6_3). URL: https://doi.org/10.1007/978-3-031-72970-6_3.
- [52] Yaofang Liu, Xiaodong Cun, Xuebo Liu, et al. “EvalCrafter: Benchmarking and Evaluating Large Video Generation Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 22139–22149. DOI: [10.1109/CVPR52733.2024.02090](https://doi.org/10.1109/CVPR52733.2024.02090). URL: <https://doi.org/10.1109/CVPR52733.2024.02090>.
- [53] Yuanxin Liu, Lei Li, Shuhuai Ren, et al. “FETV: A Benchmark for Fine-Grained Evaluation of Open-Domain Text-to-Video Generation”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, et al. 2023. URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/c481049f7410f38e788f67c171c64ad5-Abstract-Datasets%5C_and%5C_Benchmarks.html.
- [54] Zeyi Liu, Arpit Bahety, and Shuran Song. “REFLECT: Summarizing Robot Experiences for Failure Explanation and Correction”. In: *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*. Ed. by Jie Tan, Marc Toussaint, and Kourosh Darvish. Vol. 229. Proceedings of Machine Learning Research. PMLR, 2023, pp. 3468–3484. URL: <https://proceedings.mlr.press/v229/liu23g.html>.
- [55] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. “EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, et al. 2023. URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets%5C_and%5C_Benchmarks.html.
- [56] Fanqing Meng, Jiaqi Liao, Xinyu Tan, et al. “Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation”. In: *CoRR abs/2410.05363* (2024). DOI: [10.48550/ARXIV.2410.05363](https://doi.org/10.48550/ARXIV.2410.05363). arXiv: [2410.05363](https://arxiv.org/abs/2410.05363). URL: <https://doi.org/10.48550/ARXIV.2410.05363>.
- [57] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, et al. “HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips”. In: *ICCV*. 2019.
- [58] Saman Motamed, Laura Culp, Kevin Swersky, et al. *Do generative video models understand physical principles?* 2025. arXiv: [2501.09038](https://arxiv.org/abs/2501.09038) [cs.CV]. URL: <https://arxiv.org/abs/2501.09038>.
- [59] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, et al. “Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration”. In: *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 2024, pp. 6892–6903. DOI: [10.1109/ICRA57147.2024.10611477](https://doi.org/10.1109/ICRA57147.2024.10611477). URL: <https://doi.org/10.1109/ICRA57147.2024.10611477>.
- [60] OpenAI. *Video generation models as world simulators*. Feb. 2024. URL: <https://openai.com/index/video-generation-models-as-world-simulators/>.
- [61] OpenAI. *GPT-5 System Card*. Accessed: 2025-10-16. Aug. 2025. URL: <https://cdn.openai.com/gpt-5-system-card.pdf>.

- [62] OpenAI. *Sora 2 System Card*. Tech. rep. System Card. OpenAI, Sept. 2025.
- [63] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. “Learning transferable visual models from natural language supervision”. In: 2021.
- [64] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. “Overcoming Language Priors in Visual Question Answering with Adversarial Regularization”. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, et al. 2018, pp. 1548–1558. URL: <https://proceedings.neurips.cc/paper/2018/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html>.
- [65] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. “SAM 2: Segment Anything in Images and Videos”. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL: <https://openreview.net/forum?id=Ha6RTeWMD0>.
- [66] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, et al. “IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning”. In: *CoRR abs/1803.07616* (2018). arXiv: 1803.07616. URL: <http://arxiv.org/abs/1803.07616>.
- [67] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, et al. “DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation”. In: 2023.
- [68] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, et al. “RoboVQA: Multimodal Long-Horizon Reasoning for Robotics”. In: *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 2024, pp. 645–652. DOI: 10.1109/ICRA57147.2024.10610216. URL: <https://doi.org/10.1109/ICRA57147.2024.10610216>.
- [69] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, et al. “TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models”. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL: <https://openreview.net/forum?id=fCi4o83Mfs>.
- [70] Modi Shi, Li Chen, Jin Chen, et al. “Is Diversity All You Need for Scalable Robotic Manipulation?” In: *CoRR abs/2507.06219* (2025). DOI: 10.48550/ARXIV.2507.06219. arXiv: 2507.06219. URL: <https://doi.org/10.48550/arXiv.2507.06219>.
- [71] Enxin Song, Wenhao Chai, Weili Xu, et al. “Video-MLU: A Massive Multi-Discipline Lecture Understanding Benchmark”. In: *CoRR abs/2504.14693* (2025). DOI: 10.48550/ARXIV.2504.14693. arXiv: 2504.14693. URL: <https://doi.org/10.48550/arXiv.2504.14693>.
- [72] Gemini Robotics Team. “Gemini Robotics: Bringing AI into the Physical World”. In: *CoRR abs/2503.20020* (2025). DOI: 10.48550/ARXIV.2503.20020. arXiv: 2503.20020. URL: <https://doi.org/10.48550/arXiv.2503.20020>.
- [73] Hansi Teng, Hongyu Jia, Lei Sun, et al. “MAGI-1: Autoregressive Video Generation at Scale”. In: *CoRR abs/2505.13211* (2025). DOI: 10.48550/ARXIV.2505.13211. arXiv: 2505.13211. URL: <https://doi.org/10.48550/arXiv.2505.13211>.
- [74] Peter Tong, Ellis Brown, Penghao Wu, et al. “Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs”. In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. Ed. by Amir Globersons, Lester Mackey, Danielle Belgrave, et al. 2024. URL: http://papers.nips.cc/paper%5C_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.
- [75] C.J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979. ISBN: 9780408709293. URL: <https://books.google.com/books?id=t-pTAAAMAAJ>.
- [76] Homer Rich Walke, Kevin Black, Tony Z. Zhao, et al. “BridgeData V2: A Dataset for Robot Learning at Scale”. In: *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*. Ed. by Jie Tan, Marc Toussaint, and Kourosh Darvish. Vol. 229. Proceedings of Machine Learning Research. PMLR, 2023, pp. 1723–1736. URL: <https://proceedings.mlr.press/v229/walke23a.html>.
- [77] Ang Wang, Baole Ai, Bin Wen, et al. “Wan: Open and Advanced Large-Scale Video Generative Models”. In: *CoRR abs/2503.20314* (2025). DOI: 10.48550/ARXIV.2503.20314. arXiv: 2503.20314. URL: <https://doi.org/10.48550/arXiv.2503.20314>.
- [78] Jing Wang, Ao Ma, Ke Cao, et al. “WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation”. In: *CoRR abs/2503.08153* (2025). DOI: 10.48550/ARXIV.2503.08153. arXiv: 2503.08153. URL: <https://doi.org/10.48550/arXiv.2503.08153>.
- [79] Peng Wang, Shuai Bai, Sinan Tan, et al. “Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution”. In: *CoRR abs/2409.12191* (2024). DOI: 10.48550/ARXIV.2409.12191. arXiv: 2409.12191. URL: <https://doi.org/10.48550/arXiv.2409.12191>.

- [80] Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. “InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency”. In: *CoRR* abs/2508.18265 (2025). DOI: [10.48550/ARXIV.2508.18265](https://doi.org/10.48550/ARXIV.2508.18265). arXiv: 2508.18265. URL: <https://doi.org/10.48550/arXiv.2508.18265>.
- [81] Xin Wang, Taein Kwon, Mahdi Rad, et al. “HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World”. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 20213–20224. DOI: [10.1109/ICCV51070.2023.01854](https://doi.org/10.1109/ICCV51070.2023.01854). URL: <https://doi.org/10.1109/ICCV51070.2023.01854>.
- [82] Yi Wang, Yinan He, Yizhuo Li, et al. “InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation”. In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: <https://openreview.net/forum?id=MLBdiWu4Fw>.
- [83] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Trans. Image Process.* 13.4 (2004), pp. 600–612. DOI: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861). URL: <https://doi.org/10.1109/TIP.2003.819861>.
- [84] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Trans. Image Process.* 13.4 (2004), pp. 600–612. DOI: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861). URL: <https://doi.org/10.1109/TIP.2003.819861>.
- [85] Zejia Weng, Xitong Yang, Ang Li, et al. “Open-VCLIP: Transforming CLIP to an Open-vocabulary Video Model via Interpolated Weight Optimization”. In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 36978–36989. URL: <https://proceedings.mlr.press/v202/weng23b.html>.
- [86] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, et al. “Video models are zero-shot learners and reasoners”. In: *CoRR* abs/2509.20328 (2025). DOI: [10.48550/ARXIV.2509.20328](https://doi.org/10.48550/ARXIV.2509.20328). arXiv: 2509.20328. URL: <https://doi.org/10.48550/arXiv.2509.20328>.
- [87] Haoning Wu. *Open Source Deep End-to-End Video Quality Assessment Toolbox*. 2022. URL: <http://github.com/timothyhtimothy/fast-vqa>.
- [88] Haoning Wu, Chaofeng Chen, Jingwen Hou, et al. “FAST-VQA: Efficient End-to-End Video Quality Assessment with Fragment Sampling”. In: *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*. Ed. by Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, et al. Vol. 13666. Lecture Notes in Computer Science. Springer, 2022, pp. 538–554. DOI: [10.1007/978-3-031-20068-7_31](https://doi.org/10.1007/978-3-031-20068-7_31). URL: https://doi.org/10.1007/978-3-031-20068-7_31.
- [89] Haoning Wu, Dongxu Li, Bei Chen, et al. “LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding”. In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. Ed. by Amir Globersons, Lester Mackey, Danielle Belgrave, et al. 2024. URL: http://papers.nips.cc/paper%5C_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets%5C_and%5C_Benchmarks%5C_Track.html.
- [90] Haoning Wu, Erli Zhang, Liang Liao, et al. “Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives”. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 20087–20097. DOI: [10.1109/ICCV51070.2023.01843](https://doi.org/10.1109/ICCV51070.2023.01843). URL: <https://doi.org/10.1109/ICCV51070.2023.01843>.
- [91] Jinbo Xing, Menghan Xia, Yong Zhang, et al. “Dyna-iCrafter: Animating Open-Domain Images with Video Diffusion Priors”. In: *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*. Ed. by Ales Leonardis, Elisa Ricci, Stefan Roth, et al. Vol. 15104. Lecture Notes in Computer Science. Springer, 2024, pp. 399–417. DOI: [10.1007/978-3-031-72952-2_23](https://doi.org/10.1007/978-3-031-72952-2_23). URL: https://doi.org/10.1007/978-3-031-72952-2_23.
- [92] An Yang, Anfeng Li, Baosong Yang, et al. “Qwen3 Technical Report”. In: *CoRR* abs/2505.09388 (2025). DOI: [10.48550/ARXIV.2505.09388](https://doi.org/10.48550/ARXIV.2505.09388). arXiv: 2505.09388. URL: <https://doi.org/10.48550/arXiv.2505.09388>.
- [93] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, et al. “Generalized Predictive Model for Autonomous Driving”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 14662–14672. DOI: [10.1109/CVPR52733.2024.01389](https://doi.org/10.1109/CVPR52733.2024.01389). URL: <https://doi.org/10.1109/CVPR52733.2024.01389>.
- [94] Lihe Yang, Bingyi Kang, Zilong Huang, et al. “Depth Anything V2”. In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. Ed. by Amir Globersons, Lester Mackey, Danielle Belgrave, et al. 2024. URL: <http://papers.nips.cc>.

- cc/paper%5C_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58 - Abstract-Conference.html.
- [95] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, et al. “CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer”. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL: <https://openreview.net/forum?id=LQzN6TRFg9>.
- [96] Shenghai Yuan, Jinfa Huang, Yongqi Xu, et al. “ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation”. In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. Ed. by Amir Globersons, Lester Mackey, Danielle Belgrave, et al. 2024. URL: http://papers.nips.cc/paper%5C_files/paper/2024/hash/25b9960c8a5bd887eb5476c951260403 - Abstract - Datasets%5C_and%5C_Benchmarks%5C_Track.html.
- [97] Kaichen Zhang, Bo Li, Peiyuan Zhang, et al. “LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models”. In: *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Association for Computational Linguistics, 2025, pp. 881–916. DOI: 10.18653/v1/2025.FINDINGS-NAACL.51. URL: <https://doi.org/10.18653/v1/2025.findings-naacl.51>.
- [98] Richard Zhang, Phillip Isola, Alexei A. Efros, et al. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068. URL: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Zhang%5C_The%5C_Unreasonable%5C_Effectiveness%5C_CVPR%5C_2018%5C_paper.html.
- [99] Dian Zheng, Ziqi Huang, Hongbo Liu, et al. “VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness”. In: *CoRR abs/2503.21755 (2025)*. DOI: 10.48550/ARXIV.2503.21755. arXiv: 2503.21755. URL: <https://doi.org/10.48550/arXiv.2503.21755>.
- [100] Junjie Zhou, Yan Shu, Bo Zhao, et al. “MLVU: Benchmarking Multi-task Long Video Understanding”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 13691–13701. DOI: 10.1109/CVPR52734.2025.01278. URL: https://openaccess.thecvf.com/content/CVPR2025/html/Zhou%5C_MLVU%5C_Benchmarking%5C_Multi-task%5C_Long%5C_Video%5C_Understanding%5C_CVPR%5C_2025%5C_paper.html.