Reference-Free Assessment of Physical Consistency in World Model-based Video Generation

Yun Oh[1] Sukmin Yun[1] *

Abstract

We introduce reference-free measures for evaluating the physical consistency of generated videos, combining relative and absolute approaches to assess fidelity. Although tools like WorldGym or WorldEval enable robotic simulation via video generation, physical fidelity gaps often prevent these environments from accurately reproducing real-world task success rates of VLA models. Unlike existing evaluation methods, which require costly human voting (Elo) or unavailable ground-truth references (FVD), our approach utilizes DROID-SLAM and SEA-RAFT to quantify physical inconsistencies, motivated by WorldScore. Videos filtered using our relative consistency assessment show an improvement in task success rates of over 8%, effectively narrowing the simulation-to-reality gap. Furthermore, our absolute assessment enables spatio-temporal localization, providing visualization of when and where physical artifacts occur.

1. Introduction

In robot learning, simulation has become a practical necessity due to the limited accessibility of real-world robotic systems. However, performance in simulation often fails to reliably transfer to real-world settings. Recent VLA models [1, 6, 7, 21] show task success rate above 95% on the LIBERO [10] benchmark. In contrast, more recent environments [4, 20] are sufficiently difficult that many existing policies struggle to achieve meaningful success rates.

Meanwhile, world model–based approaches [9, 13] have shown promising alignment with real-world outcomes, suggesting a potential path toward more realistic evaluation. Nevertheless, these approaches introduce a new challenge: the generated visual rollouts often suffer from physical inconsistencies, such as object morphing, making it difficult to reliably assess task success. Standard metrics for videos like FVD [16] fall short here, as they measure statistical distribution rather than the pixel-level physical integrity required for robotic manipulation.

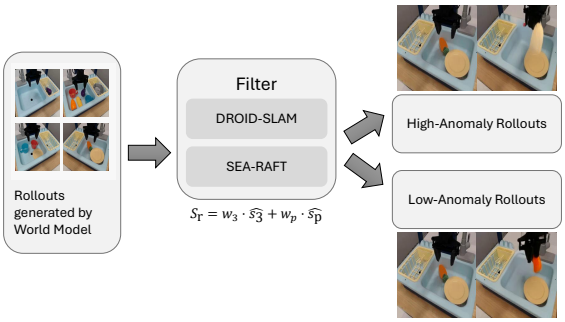

Figure 1. **Overview of Relative Assessment**. Generated rollouts are evaluated using DROID-SLAM [15] and SEA-RAFT [18] to obtain relative anomaly scores, which enable filtering and ranking of samples based on physical plausibility.

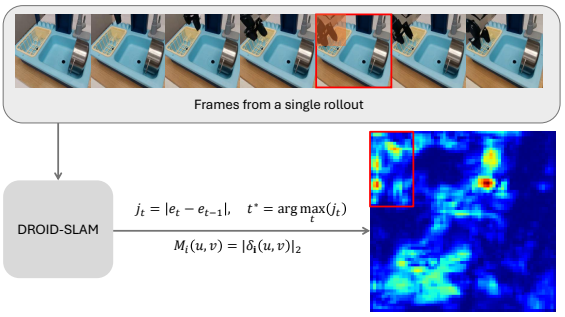

Figure 2. **Overview of Absolute Assessment**. Frames from a single rollout are evaluated using DROID-SLAM [15] to obtain the anomaly heatmap, which shows where and when physical feasibility collapses.

Although obvious artifacts can be easily identified, evaluating the overall quality of generated rollouts becomes challenging when multiple imperfections are present. In such cases, it is unclear which samples are more suitable for evaluation. VLA models require high temporal coherence to maintain action continuity during task execution, so our method adopts several strategies from the recent world model benchmarks [2, 8].

In this paper, we introduce relative and absolute assessments of generated rollouts: First, we introduce a relative assessment for quantitative screening. By comparing the 3D

*Correspondence to: sukminyun@hanyang.ac.kr
[1]Hanyang University ERICA

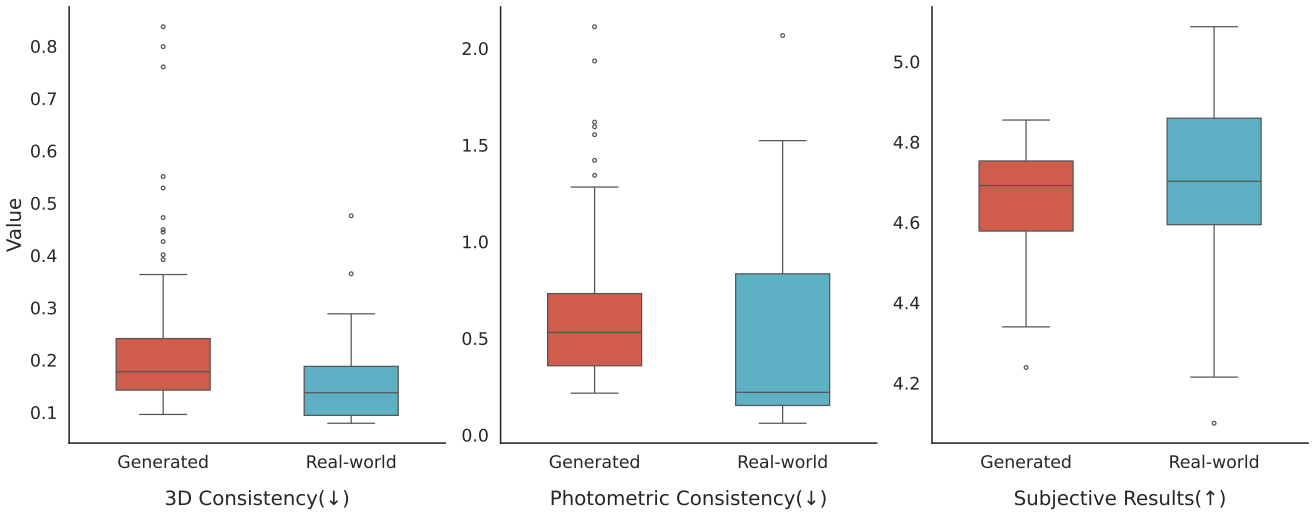

Figure 3. **Quality Scores in Real and Generated Videos**. 3D and photometric consistency exhibit distinctive gaps between real and generated distributions, whereas median of subjective quality remains comparable. Generated videos contain a significantly higher number of outliers in 3D and photometric consistency.

and photometric consistency of generated rollouts against real-world videos, we derive an 'anomaly score' to filter out physically implausible sequences. This pre-processing ensures that the subsequent VLM reward model evaluates only high-fidelity rollouts, preventing the policy from exploiting model artifacts. Second, we propose an absolute assessment mechanism for fine-grained qualitative diagnosis. Beyond simple filtering, we pinpoint when and where consistency collapses by employing DROID-SLAM [15]'s factor graphs. This generates pixel-wise heatmaps that visualize the world model's failure modes.

Overall, our work empirically improves prior methods by: (1) narrowing the focus to object morphing via a combination of entropy-based 3D analysis and photometric consistency, (2) introducing a visualization derived from geometric shift computations to point out where and when the most drastic physical failure occurs, and (3) demonstrating that our filtering mechanism improves the task success rate of VLA models by over 8% by effectively narrowing the simulation-to-reality gap.

2. Methods

Among the four quality metrics in WorldScore [2], we exclude style consistency and subjective quality due to subtle variance. Then we select 3D consistency and photometric consistency, since they indicate significant statistical gaps between real and generated videos as shown in Fig. 3. Accordingly, we devise relative assessment as Shannon entropy-weighted sums of consistencies, and absolute assessment based on pixel distances of adjacent frames.

2.1. Relative Assessment: Quantitative Screening

For 3D consistency, we calculate the reprojection error across co-visible pixels between successive frames, utilizing DROID-SLAM [15]. In terms of photometric consistency, we compute the Average End-Point Error(AEPE), using SEA-RAFT [18]. Both scores are computed using the publicly available source code from WorldScore [2].

While both metrics are fundamentally significant, we utilize Shannon entropy to objectively evaluate their respective informativeness and determine the optimal weighting for each component. To evaluate the informativeness of each consistency score for anomaly detection, we calculate the Shannon entropy:

$$H(p) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

Accordingly, we calculate weights using the stable softmax function as below:

$$w_k = \frac{\exp(-H_k)}{\sum_j \exp(-H_j)}$$

Then we assign each weight to the metrics in question.

Subsequently, we obtain relative anomaly scores with the weighted sum:

$$S_{\mathrm{r}} = w_3 \cdot \hat{s}_3 + w_p \cdot \hat{s}_{\mathrm{p}}$$

where $\hat{s}_k = (s_k - \mu_k)/\sigma_k$ denotes the z-score normalized score of the metric k. The stability-based weighting mechanism ensures that the anomaly detection is robust even when one metric exhibits high variance due to generation noise.

Category	Task	# Trials	WorldGym # Successes	Real-world # Successes
Visual gen	Put Eggplant into Pot (Easy Version)	10	7	10
Motion gen	Lift Eggplant	10	7.5	7.5
Physical gen	Put Carrot On Plate	10	4	8
Semantic gen	Stack Blue Cup on Pink Cup	10	6	4.5
Language Grounding	Put {Blue Cup, Pink Cup} on Plate	10	8.5	9.5
		Mean Success Rate	**66.0** $\pm$ 6.6%	**79.0** $\pm$ 5.5%

Table 1. **Baseline Task Performance from Prior Work**. Task success numbers are taken directly from WorldGym [13]. Average success rates $\pm$ StdErr are computed across 50 total rollouts, following OpenVLA calculated across 170 total rollouts for 17 tasks.

2.2. Absolute Assessment: Qualitative Analysis

When there are not enough samples, we can borrow relative anomaly scores from real-world recordings in similar environments. However, since the score alone is not sufficient to evaluate overall quality, we propose a two-step analysis based on DROID-SLAM [15] factor graph, focusing on 3D consistency to identify specific failure points.

We specifically prioritize 3D consistency for absolute assessment and qualitative visualization. This choice is justified by its lower Shannon entropy compared to photometric consistency, indicating higher informational certainty. Furthermore, the reprojection error in 3D consistency provides a more direct physical grounding, enabling us to precisely localize where and when physical anomalies occur within the generated sequence.

2.2.1. Temporal Detection

We first monitor the reprojection error e_t from the factor graph updates. To exclude the initial warmup, we only consider frames where $t > 7$, following the original implementation. The frame t^* with the maximum geometric shift is then identified by the error jump j_t:

$$j_t = |e_t - e_{t-1}|, \quad t^* = \arg\max_t(j_t)$$

2.2.2. Spatial Localization

Within the selected frame t^*, we examine the residual flow $\delta_i \in \mathbb{R}^{H\times W\times 2}$ for each edge i on the factor graph, where H and W denote the height and width of the frame. Since each δ_i is a dense vector field which contains error values for all pixels, we calculate the pixel-wise magnitude $M_i(u, v)$ to visualize the intensity of distortions:

$$M_i(u, v) = \|\delta_i(u, v)\|_2$$

where (u, v) represents the coordinates of the pixels.

To select the most representative artifact, we calculate the mean magnitude S_i for each edge:

$$S_i = \frac{1}{HW}\sum_{u,v} M_i(u, v), \quad i^* = \arg\max_i(S_i)$$

3. Experiments

We specifically focus our evaluation on OpenVLA [5] due to the comprehensive accessibility of real-world out-of-distribution(OOD) of BridgeData V2 [17] with WidowX robot experiment videos and success rates. While our methodology applies to any robot policy evaluation, this allows a quantitative validation of our proposed measures by comparing generated low-anomaly rollouts against established real-world performance.

5 tasks are selected based on the availability of real-world experiment videos on the OpenVLA website, with one task representing each of the categories suggested in their study. The details of the five tasks are summarized in Tab. 1. Then 30 videos per task are generated employing WorldGym [13] with a random seed and OpenVLA.

3.1. Relative Assessment

To sort high and low anomaly rollouts, we generate 30 videos for each of the five tasks, resulting in a total of 150 videos. To ensure variability, each video was produced using a unique random seed, starting from 42 and incrementing for each subsequent generation.

Both real-world and generated videos' scores are put into equal-width bins(n=30), where $p(x_i)$ denotes the frequency of scores falling into the i-th bin. The number of bins was chosen to align with the sample size of our real-world dataset. Analysis across these datasets shows that 3D consistency has lower entropy compared to photometric consistency, which can be interpreted as 3D consistency having higher informational certainty and representational stability. Specific numbers are described in Tab. 2.

The distribution of relative anomaly scores Fig. 4 shows a noticeable shift, where generated videos tend to have higher scores and wider variance compared to real-world videos. This suggests that generated videos with low anomaly scores may better represent the real-world capabilities of VLA models.

To evaluate the generated rollouts, GPT-4o [12] was employed as a reward model. The evaluation follows the ap-

Metric	Property	Generated	Real-world
3D Consist	Entropy	2.16	1.73
	Weight	**0.67**	**0.65**
Photo Consist	Entropy	2.88	2.35
	Weight	0.33	0.35

Table 2. **Shannon Entropy and Weights of Real and Generated Videos**. The weights derived from generated and real-world datasets align, even with disparity between two datasets.

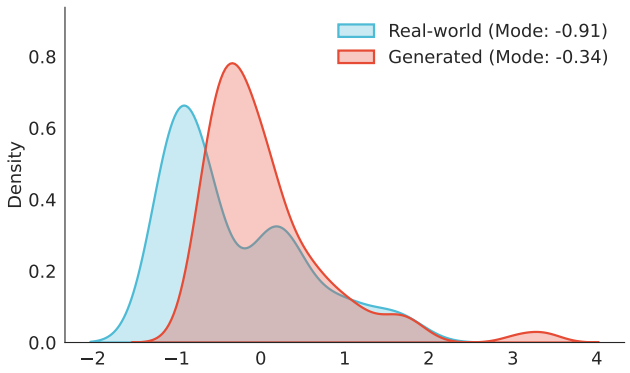

Figure 4. **Distribution of Relative Anomaly Scores**. Lower score is better. Density allows normalized comparison for varied OpenVLA rollouts across unequal sizes.

proach described in WorldGym [13]. Rollouts per task were categorized into high-anomaly and low-anomaly groups (n=10 each) based on their relative assessment scores, while excluding the middle 10 rollouts to ensure better distinction between the two groups. Additionally, as a baseline for comparison, we evaluate 10 randomly generated rollouts without fixed seeds or anomaly-based filtering. The result is described in Fig. 5

Applying relative anomaly scores for filtering improves evaluation accuracy and helps bridge the simulation-to-reality gap. However, we observe a specific limitation in tasks like 'Lift Eggplant', where low-anomaly rollouts fail to capture the real-world VLA ability in Tab. 3. This is because when the robot fails to contact the object entirely, the lack of visual distortion results in misleadingly low scores. This confirms that physical consistency is a necessary but insufficient condition for semantic success, which we leave for future studies.

3.2. Absolute Assessment

The rollouts used are initialized from a frame where the robot arm is outside the field of view. Consequently, the world model lacks sufficient visual grounding to accurately reconstruct the arm's morphology, leading to inconsistent or hallucinated representations in the generated sequence.

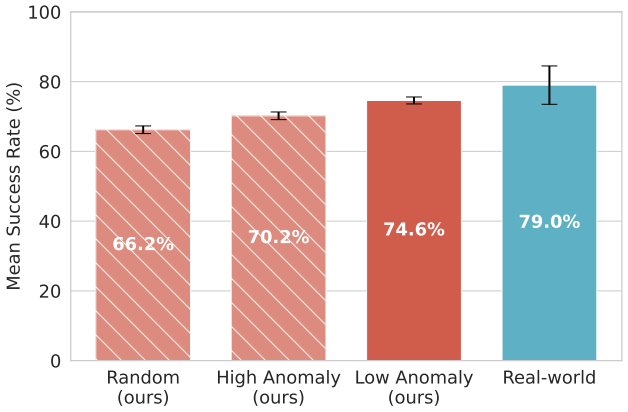

Figure 5. **Success Rates across different conditions** after filtering with our relative assessment with the real-world result [5]. Our results show that the low-anomaly group achieves success rates closest to those of real-world videos. We repeat GPT-4o evaluation 10 times per video to reduce evaluator variance, resulting in a smaller SE compared to real-robot experiments. Detailed per-task results are provided in Tab. 3

The resulting magnitude map M_{i^*} is used as our Artifact Heatmap. This highlights specific regions where the generative model fails to maintain 3D consistency.

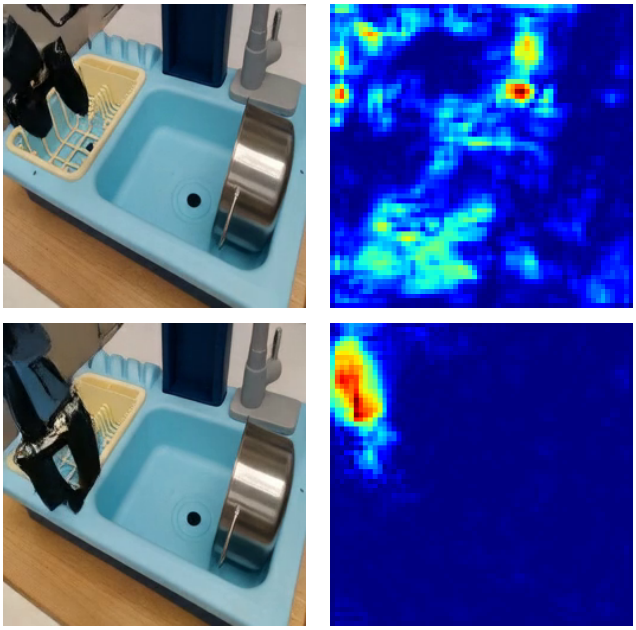

Figure 6. **Artifact Heatmap** shows the result of the absolute assessment. Red colors imply low fidelity of the specific edges. The first row displays a deformed robot arm and the second row shows a thickened robot arm, both of which are clearly characterized in the adjacent heatmap.

Task	Random (ours)	High Anomaly (ours)	Low Anomaly (ours)	Real-world [5]
Put Eggplant into Pot (Easy Version)	70	86	73	100
Lift Eggplant	74	84	64	75
Put Carrot On Plate	48	39	75	80
Stack Blue Cup on Pink Cup	63	72	79	45
Put {Blue Cup, Pink Cup} on Plate	76	70	82	95
Mean Success Rate	**66.2** $\pm$ 1.1%	**70.2** $\pm$ 1.1%	**74.6** $\pm$ 1.0%	**79.0** $\pm$ 5.5%

Table 3. **Task Performance (%)** shows relative assessment results. We select 5 tasks in 17 OOD tasks from OpenVLA [5]. Each task is selected to follow five categories as we elaborate earlier Sec. 3. Random, High Anomaly, Low Anomaly group data is from our experiments, which conduct 10 sets of 10 trials (100 iterations) per task. Real-world data is sourced from OpenVLA [5] as well as Tab. 1, which reports results based on 10 trials per task. Among three groups of generated rollouts, Low Anomaly group by our metric yield a mean success rate that approximates Real-world performance.

3.3. Challenges in Cross-Model Generalization

To stress-test the robustness of this metric across diverse distributions, we extended our evaluation to generated clips from representative models [11, 14, 19].

Sora. The clip utilized in our analysis is the 'woman walking down a Tokyo street' sourced from the official OpenAI website. While the video exhibits synthetic artifacts easily recognizable by humans, these anomalies remain consistent throughout the sequence. Consequently, the 3D consistency score does not exhibit significant fluctuations, rendering the analysis of a single captured frame unreliable. Indeed, the sequence recorded a Mean Score of 1.9326 with a Max Jump of only 0.0063, confirming the high temporal persistence of artifacts.

Figure 7. **Sora**. The left image shows the original frame, and the heatmap is overlaid on the original image. This is not a relatively implausible frame to human perspective.

Grok. We design a prompt describing a hand grasping an apple that exhibits surreal, gelatinous deformation instead of rigid-body physics. In this case, the frame in question was detected, but the heatmap was distributed broadly over the sequence, lacking the spatial precision to isolate the point of consistency collapse. The sequence recorded a Mean Score of 1.6282 with a Max Jump of 0.0041.

LucyEdit. Because LucyEdit often produces artifacts from the frame, our detector paradoxically identifies the 'improved' frame as an anomaly. Although it successfully captures a deviation in the sequence, it fails to isolate physical failure due to the absence of a clean reference frame. The input natural video recorded a mean score of 0.1874 with a max jump of 0.0039 and the output edited footage showed a mean score of 0.1432 with a max jump of 0.0038, which demonstrates certain generated footages are more consistent than natural videos according to the reprojection error.

Figure 8. **Grok-generated physically implausible sequence and the heatmap**. The left image shows the original frame, and the heatmap is overlaid on the original image. The heatmap shows the limitation in our method to capture specific location where the physical inconsistency is detected in the frame.

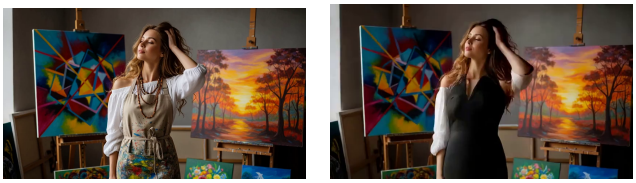

Figure 9. **LucyEdit** original sample's first frame (left) and the edited footage's first frame (right). In the edited frame, a distinct 'V' shape artifact is visible on her chest around the black garment. This contour is unnatural and does not correspond to either clothing drape or anatomy.

Noise-corrupted videos. Stable Diffusion 3 [3] is chosen to create an anomaly frame. We use a custom recorded footage, and we also use generated videos from the three models mentioned above.

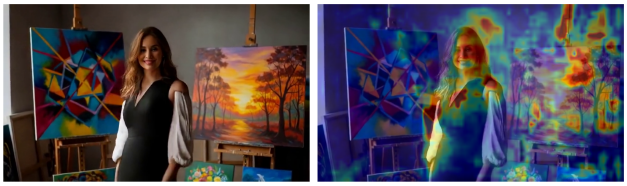

Figure 10. The edited sequence from **LucyEdit** displays more significant artifacts in the preceding frames compared to this sample. The specific editing instruction provided to the model was: 'Change her clothes to a black mini dress for funeral'.

Figure 11. **Noise-corrupted**. These are one of samples, which is a custom recorded footage. The left one is the original frame, and the right one is the corrupted image that replaced the original frame for the experiments. Both the heatmap generation and score computation fail due to zero values encountered during the calculation.

With several versions of augmentation applying various prompts and hyperparameter tuning, our method always returns undesired frames that are not corrupted. In sequences containing augmented frames, the reprojection error occasionally drops to zero, which likely indicates a failure to establish valid pixel correspondences. Standard video sequences, whether generated or natural, consistently yield non-zero scores, reflecting a continuous flow of identifiable features.

This cross-model analysis identifies two fundamental challenges for current SLAM-based assessment: (1) the collapse of tracking under severe structural inconsistencies where correspondences cannot be established, and (2) localization instability under non-rigid deformations or dynamic lighting. Addressing these challenges remains an important direction for future work.

4. Discussion

Conclusion. We present a practical application of existing consistency metrics for analyzing generated sequences, specifically within the robotic simulation domain. This approach does not require a ground truth, which is often unavailable when assessing outputs from modern generative models. In scenarios where ground-truth references are unavailable, our metrics serve as a preliminary filter to ensure basic physical consistency. By narrowing the simulation-to-reality gap in VLA model evaluation, this approach offers a scalable alternative to manual inspection for verifying robot-centric visual rollouts. By extension, our relative anomaly score can serve as a reward signal for refining world models.

Limitation. Our assessment relies on quality metrics devised for static generation, which can fail under severe structural inconsistencies, producing zero values that limit applicability to certain generation models. Anomaly scores can also be misleadingly low when task failure involves no physical contact. Beyond these metric-level limitations, our evaluation is primarily conducted on OpenVLA, leaving generalization to other VLA models unexplored. Future work will extend evaluation across diverse VLA and world models, and investigate semantic or contact-aware signals to address these limitations.

Acknowledgement

We thank the authors of WorldGym and WorldScore for providing the open-source codebases that served as the foundation for our experiments. We are also grateful to the OpenVLA contributors for their real-world experiment rollouts, which were essential for the execution of our comparative analysis.

References

[1] Kevin Black, Noah Brown, et al. π_0: A vision-language-action flow model for general robot control, 2026. 1

[2] Haoyi Duan, Hong-Xing Yu, et al. Worldscore: A unified evaluation benchmark for world generation, 2025. 1, 2

[3] Patrick Esser, Sumith Kulal, et al. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 5

[4] Songhao Han, Boxiang Qiu, et al. Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation, 2025. 1

[5] Moo Jin Kim, Karl Pertsch, et al. Openvla: An open-source vision-language-action model, 2024. 3, 4, 5

[6] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025. 1

[7] Moo Jin Kim, Yihuai Gao, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning, 2026. 1

[8] Dacheng Li, Yunhao Fang, et al. Worldmodelbench: Judging video generation models as world models, 2025. 1

[9] Yaxuan Li, Yichen Zhu, et al. Worldeval: World model as real-world robot policies evaluator, 2025. 1

[10] Bo Liu, Yifeng Zhu, et al. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. 1

[11] OpenAI. Sora: Creating video from text. https://openai.com/index/sora/, 2024. Accessed: 2026-04-18. 5

[12] OpenAI, :, et al. Gpt-4o system card, 2024. 3

[13] Julian Quevedo, Ansh Kumar Sharma, et al. Worldgym: World model as an environment for policy evaluation, 2025. 1, 3, 4

[14] DecartAI Team. Lucy edit: Open-weight text-guided video editing, 2025. 5

[15] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras, 2022. 1, 2, 3

[16] Thomas Unterthiner, Sjoerd van Steenkiste, et al. Towards accurate generative models of video: A new metric & challenges, 2019. 1

[17] Homer Walke, Kevin Black, et al. Bridgedata v2: A dataset for robot learning at scale, 2024. 3

[18] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow, 2024. 1, 2

[19] xAI. Grok imagine. https://x.ai/news/grok-imagine-api, 2026. Accessed: 2026-04-18. 5

[20] Shiduo Zhang, Zhe Xu, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks, 2024. 1

[21] Jinliang Zheng, Jianxiong Li, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model, 2025. 1