WORLDPREDICTION: A Benchmark for High-level World Modeling and Long-horizon Procedural Planning

Delong Chen [*] [1] [2] **Willy Chung** [*] [1] [3] **Yejin Bang** [1] [2] **Ziwei Ji** [1] [2] **Pascale Fung** [1] [2]

Abstract

World models predict future world states resulting from actions, enabling AI agents to perform planning in diverse environments. We introduce WorldPrediction, a video-based benchmark for evaluating world modeling and procedural planning capabilities of different models. In contrast to prior works that focus primarily on low-level world modeling and robotic motion planning, WORLDPREDICTION is the first benchmark that emphasizes actions with temporal and semantic abstraction. Given initial and final world states, the task is to distinguish the proper action (WORLDPREDICTION-WM) or the properly ordered sequence of actions (WORLDPREDICTION-PP) from a set of counterfactual distractors. As such, to prevent models from exploiting low-level continuity cues in background scenes, we provide "action equivalents" – identical actions observed in different contexts – as candidates for selection. This benchmark is grounded in a formal framework of partially observable semi-MDP, which ensures better reliability and robustness of the evaluation. We conduct extensive human filtering and validation on our benchmark and show that current frontier models barely achieves 57% accuracy on High-level World Modeling and 38% on Long-horizon Procedural Planning whereas humans are able to perfectly solve both tasks.

1. Introduction

Advanced machine intelligence relies critically on two foundational capabilities: world modeling and procedural planning (LeCun, 2022; Ha & Schmidhuber, 2018). World modeling allows agents to internally simulate future world states, enabling them to optimize their actions without trial-and-error in the real world or reliance on explicit reward signals. Procedural planning (Chang et al., 2020) involves determining ordered sequences of actions to achieve long-horizon goals. These capabilities represent key steps toward developing AI systems that can reason effectively, act responsibly, and interact smartly with complex environments.

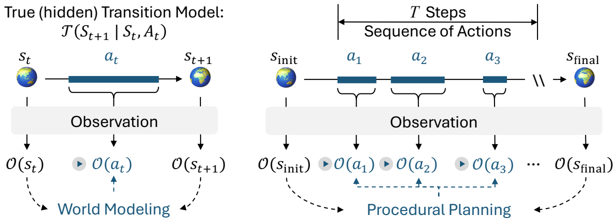


Figure 1. **Theoretical formulation of WorldPrediction**. Latent world states (s) and high-level actions (a) evolve according to a hidden transition model $\mathcal{T}$, which is not directly accessible. Instead, an observation model $\mathcal{O}$ maps these latent variables into visual observations, producing images $\mathcal{O}(s)$ depicting states and video segments $\mathcal{O}(a)$ depicting actions.

Recent advances in low-level world modeling and planning have achieved significant progress in intuitive physics understanding (Garrido et al., 2025), robotic motion control (Zhou et al., 2024a), navigation (Koh et al., 2021; Bar et al., 2024) and autonomous driving (Wang et al., 2024b). These scenarios typically involve precise physical dynamics and high-frequency control without any semantic or temporal abstraction. However, skilled human activities require reasoning at a higher level, where individual actions span longer, non-uniform durations and encapsulate multiple lower-level primitive actions (Sutton et al., 1999). Abstraction enables efficient long-horizon reasoning in complex tasks by significantly condensing the exponentially growing search space. It can also reduce the sensitivity to low-level variations and thereby enhancing generalization. Moreover, it aligns closely with human cognition, improving interpretability and facilitating better communication.

We propose WorldPrediction, a benchmark for evaluating high-level world modeling and long-horizon procedural planning. It consists of two sub-benchmarks:

[*] Equal contribution [1] Meta FAIR Paris [2] The Hong Kong University of Science and Technology [3] Sorbonne Université, CNRS, ISIR. Correspondence to: Delong Chen <delong.chen@connect.ust.hk>, Pascale Fung <pascale-fung@meta.com>.

ICML 2025 Workshop on Assessing World Models. Copyright 2025 by the author(s).

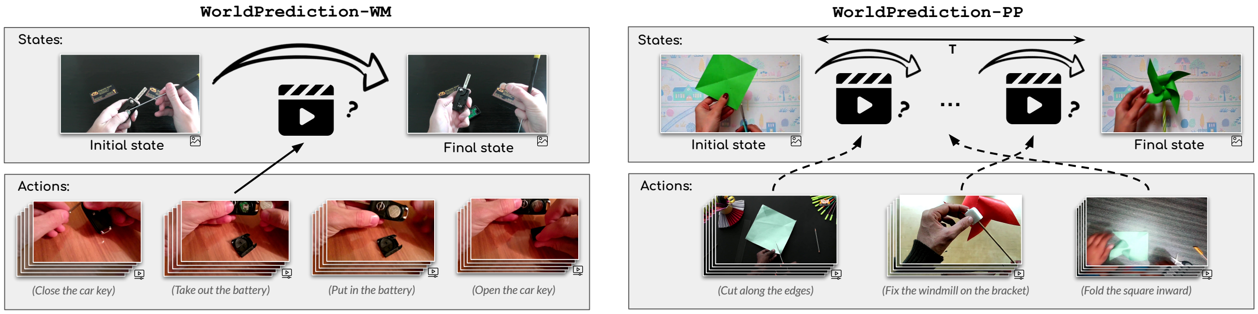


Figure 2. WORLDPREDICTION-WM and WORLDPREDICTION-PP task formulation. For World Modeling, the objective is to select which action clip depicts the transition from initial to final state. For Procedural Planning, the objective is to select which sequence of action clips ($T \in [3, 10]$) is correctly ordered to depict the transition from initial to final state. *(The actual samples in the benchmark do not contain any text, here the actions are annotated for vizualisation purposes)*

WORLDPREDICTION-WM assesses whether the model understands the causalities of semantically and temporally abstract actions in real-world skilled human activities; WORLDPREDICTION-PP further extend the evaluation to procedural planning over extended temporal horizons, in contrast to existing benchmarks that typically focus on short span of only 3-4 steps (Chang et al., 2020). Key features of the WORLDPREDICTION benchmark include:

1) Diverse Actions and Tasks. The benchmark covers a broad spectrum of human activities, such as cooking, household repair, technical maintenance, furniture assembly, health care, etc. Samples are sourced from five datasets encompassing instructional web videos as well as egocentric and exocentric recordings of skilled human activities.

2) Discriminative Formulation. The benchmark adopts a multiple-choice task formulation, where models select correct action or action sequences from a set of distractors. It facilitates direct comparisons between diverse world model / planner architectures (*e.g.*, predictive vs. generative), modality representations (*e.g.*, VLMs vs. diffusion).

3) Shortcut Mitigation. The benchmark represents states and actions using observations. To ensure that the WORLDPREDICTION evaluates the understanding of action-state causality and to discourage models from exploiting superficial continuity cues, we provide "**action equivalents**": identical actions captured in different backgrounds or observed from different viewpoints as action candidates.

We establish baseline performance on WORLDPREDICTION using several SOTA approaches, including VLMs, Socratic LLMs, video diffusion models, and Open-Event Procedural Planning (OEPP) models (Wu et al., 2024). Overall results on WORLDPREDICTION demonstrate that while better perception on larger models yield expected improvements, a substantial gap still remains between the highest-performing model (57.0% on WORLDPREDICTION-WM and 38.1% on WORLDPREDICTION-PP) and human performance.

2. Related Works

World Modeling. World Modeling is a fundamental capability of autonomous intelligent systems (LeCun, 2022), which consists in leveraging an internal representation of the world to predict and understand how the state of the world evolves under different actions. Current works can be separated into *predictive* and *generative* world modeling. Models such as I-JEPA (Assran et al., 2023), V-JEPA (Bardes et al., 2024) or DINO-WM (Zhou et al., 2024a) aim to predict latent representations of future states of the world, which has also been shown to develop a basic understanding of intuitive physics (Garrido et al., 2025). On the other hand, *generative* world modeling uses denoising backbones to generate the next world state, a capability better fit for exploration and simulation of the real world as shown by Genie (Bruce et al., 2024) or UniSim (Yang et al., 2023). Due to the complexity of the real world, current world models have been explored either in synthetic environments (Kim et al., 2023; Hafner et al., 2023; Garrido et al., 2024; Gupta et al., 2024), or in real world environments with relatively constrained action spaces such as *robotics* (Hafner et al., 2019; Zhou et al., 2024b; Wu et al., 2023; Mendonca et al., 2023) with manipulation-based actions, *autonomous driving* (Guan et al., 2024; Hu et al., 2023; Wang et al., 2024a;b) with vehicle control actions, and *navigation* (Shah et al., 2023; Koh et al., 2021; Bar et al., 2024) with spatial movement actions.

Procedural Planning. Given an initial and final state, Procedural Planning refers to the ability of predicting a sequence of actions which would bring the initial state to the final state. While that formulation is common in robotic control (Sun et al., 2022; Lynch et al., 2023) for low-level manipulation tasks, in this work we focus on human-centric procedural planning with higher-level actions (e.g., "remove the battery", "attach a table leg") (Ben-Shabat et al., 2021; Damen et al., 2022), mostly from instructional videos (Chang et al., 2020; Tang et al., 2019; Zhukov et al., 2019), which inherently involves deeper semantic reasoning and abstraction of granular actions. Recent work to tackle procedural planning include using weak supervision from language (Zhao et al., 2022), encoding intermediate state transitions (Niu et al., 2024) or using a condensed action state learning method (Li et al., 2023) Alternatively, text-supervision using LLMs have shown to be effective (Liu et al., 2023; Wang et al., 2023a; Islam et al., 2024), as well as diffusion-based approaches (Wang et al., 2023b). Despite recent attempts at expanding the scope of procedural planning (Wu et al., 2024; Patel et al., 2023), the evaluation of the task is still over-reliant on human annotated text labels of actions to convey interpretable plans, which motivates the formulation of WorldPrediction.

3. The WorldPrediction Benchmark

Task Formulation Given two images representing respectively an initial and final state, the objective of WorldPrediction is to select the correct transition that happened as shown in Fig 2. In WorldPrediction-WM, the candidates are singular video segments depicting various actions, while in WorldPrediction-PP, the candidates are shuffled sequences or video segments representing a plan of varying length. All actions are replaced by action equivalents, detailed below. We provide an extensive formal grounding of WorldPrediction in Appendix B and data statistics in Appendix C

Data Source We use the official dataset splits for evaluation of 5 different datasets: COIN (Tang et al., 2019), CrossTask (Zhukov et al., 2019), EgoExo4D (Grauman et al., 2024), EPIC-KITCHEN-100 (Damen et al., 2022), and IKEA-ASM (Ben-Shabat et al., 2021). The test split for COIN and validation splits for CrossTask, EPIC-KITCHENS-100, EgoExo4D, and IKEA-ASM. For WorldPrediction-PP, we use number of action steps $T \in \{3, 4\}$ for COIN and CrossTask, and $T \in \{3, \dots 10\}$ for the remaining. The action sequence are sampled using a sliding window following previous works.

Distractor Sampling. To rigorously test action discrimination, each correct action is presented alongside three distractors, resulting in four total candidates. For WorldPrediction-WM, distractors are plausible alternative actions drawn from the same task context (*i.e.,* same video) but incompatible with the observed state transition. For WorldPrediction-PP, distractors are generated by shuffling the action sequences, preserving action-level plausibility while disrupting temporal correctness.

World Model / Planner		WorldPrediction -WM	WorldPrediction -PP
VLMs	InternVL2.5 (2B)	20.0	21.05
	InternVL2.5 (4B)	29.8	27.9
	InternVL2.5 (26B)	30.2	30.0
	InternVL2.5 (38B)	50.3	31.1
	Qwen2.5-VL (3B)	21.6	29.1
	Qwen2.5-VL (7B)	45.5	32.5
	Qwen2.5-VL (32B)	49.0	33.5
	Qwen2.5-VL (72B)	**57.0**	**36.7**
Socratic LLMs	Llama-3.1 (8B)	48.7	26.7
	Llama-3.1 (70B)	49.8	31.2
	Llama-3.3 (70B)	52.2	35.1
	Llama-4 Scout	52.7	32.8
	Llama-4 Maverick	53.6	34.7
	Qwen2.5 (3B)	44.0	25.6
	Qwen2.5 (7B)	49.1	28.4
	Qwen2.5 (32B)	39.2	29.1
	Qwen2.5 (72B)	48.5	30.7
	DeepSeek-R1 (distilled)	50.8	28.4
	Gemini-2.0	**55.6**	33.5
	GPT-4o	52.0	33.7
	Claude-3.5-sonnet	53.3	**38.1**
Video Diffusion	I2VGenXL	26.1	N/A
	I2VGenXL + DINOv2	26.7	
	CogVideoX	30.1	
	CogVideoX + DINOv2	30.5	
OEPP	MLP	N/A	36.8
	Transformer		34.2
	PDPP		34.4

Table 1. Performance comparison on WorldPrediction-WM and WorldPrediction-PP accuracy (%).

Action Equivalent Retrieval. To mitigate shortcut learning from low-level visual continuity cues, we employ *action equivalents*: visually different yet semantically identical actions captured in alternate backgrounds or viewpoints (more details in B.3). For COIN, CrossTask, EPIC-KITCHENS-100, and IKEA-ASM, actions sharing the same textual label constitute equivalents. For EgoExo4D, where explicit temporal boundaries are unavailable, we segment actions by computing midpoints between consecutive timestamps and discard segments shorter than 5 seconds. We select the egocentric view for actions to clearly observe detailed hand movements and use exocentric viewpoints for state observations due to their comprehensive scene coverage.

Sample Filtering. To filter out nosiy observations, we compute distances between visual features of initial and final states using pretrained visual embeddings (DINOv2 (Oquab et al., 2024)). Samples exceeding predefined thresholds (2.75 for WorldPrediction-WM, 10 for WorldPrediction-PP) are excluded due to excessively drastic or incoherent scene transitions. For EgoExo4D, we additionally remove samples in which critical task-relevant visual information is obstructed by the human subject by prompting a VLM (more details in D). We further removed samples where there are too little difference between its initial and final states. These samples usually corresponded to a static segment in instructional videos, or only slight body movement in EgoExo4D videos (as shown in Fig. 7).

4. Evaluation Results

Models. We establish initial baseline performance on WORLDPREDICTION using VLMs, Socratic LLMs, video diffusion models, and Open-Event Procedural Planning (OEPP) models. VLMs and Socratic LLMs serve as both world models and procedural planners due to their flexibility, while diffusion is tailored to world modeling and OEPP is only for planning. These baselines cover both widely evaluated open-source models as well as closed-source frontier models, serving primarily to provide initial reference points for future research. Additional model information are in Appendix E.1

Performance Comparison. Table 1 summarizes model performances on the WORLDPREDICTION benchmark. In the WORLDPREDICTION-WM task, smaller-scale VLMs perform near random chance levels, with InternVL2.5 (4B) and Qwen2.5-VL (3B) model notably struggling to produce outputs that choose from given options. There is a significant breakthrough in world modeling performance past a certain model scale, with a jump of roughly 20% from 26B to 38B for InternVL2.5, and from 3B to 7B for Qwen2.5-VL. However, it is interesting to note that long-horizon procedural planning do not show a significant boost in performance with model size. Socratic LLMs,using high-quality captions generated by Qwen2.5-VL (72B), achieve comparable results to VLMs. The best performing LLMs are the closed-source Gemini-2.0 for world modeling at 55.6% and Claude-3.5 for procedural planning at 38.1%. Interestingly for socratic LLMs, the best performing model at world modeling does not translate to the best one in procedural planning, we hypothesize that perception is an important component for model to be able to extend their single-step performance to longer-horizon tasks. Additionally, it can be interpreted as a trade-off between stronger reasoning capabilities without visual grounding using socratic LLMs, and better perceptual grounding using VLMs but no explicit reasoning.

Video diffusion models exhibit comparatively lower performance, with CogVideoX-I2V reaching 30.1% and I2VGenXL achieving 26.1%. These results suggest pixel-space generation struggles to effectively capture detailed action-state causal relationships (diffusion generation are shown in D and Fig. 9), and that using better image features (DINOv2 features instead of RGB) for candidate selection does not have much impact on the results. Another limitation of diffusion models is the absence of a reliable method for selecting the correct candidate sequence. Although using the final frames may appear intuitive, it proves ineffective in accurately linking the transition to $\mathcal{O}(s_{t+1})$ For the WORLDPREDICTION-PP task, OEPP-based planners perform at a comparable level with the best zero-shot large models' performance, while being significantly smaller.

Human Evaluation and Filtering. To ensure the quality and robustness of the WORLDPREDICTION benchmark, we conducted a large-scale human evaluation and filtering process. We initially constructed 1,500 samples for both the World Modeling and Procedural Planning tasks. Each sample was then independently solved by two different annotators, following detailed task-specific instructions and solved example demonstrations, with a total of **80** different human annotators to annotate all 3000 samples. We adopted a conservative filtering criterion: only samples where both annotators independently provided the correct answer were retained. After filtering, we obtained **825** high-quality samples for WORLDPREDICTION-WM and **570** samples for WORLDPREDICTION-PP, ensuring effectively perfect human performance on WorldPrediction. Notably, due to the increased complexity of the Procedural Planning task — which requires reasoning over temporally extended sequences rather than single transitions — a smaller proportion of samples were kept. These human evaluation results highlight the difficulty of our benchmark: the current best models on WORLDPREDICTION-WM, Gemini-2.0 and Qwen2.5, achieve only 55-57% accuracy, with most models ranging between 40-50% accuracy as shown in Table 1. For WORLDPREDICTION-PP, even trained planners such as OEPP reach only around 35% accuracy, and zero-shot frontier models around 37%, highlighting a significant gap between machine and human performance, especially for procedural planning where compounding errors at longer horizon are inevitable for current models. Further details regarding the annotation process, including inter-annotator agreement scores, annotation instructions, and annotator workload distribution are provided in Appendix A.

5. Conclusion

In this work, we introduced WorldPrediction, the first benchmark designed to assess high-level world modeling and long-horizon procedural planning from purely visual observations. Unlike prior efforts that focused on low-level physical dynamics or short-horizon tasks, WORLDPREDICTION emphasizes semantic and temporal abstraction, better aligning with the properties of understanding high-level human activities. Evaluations across SOTA VLMs, LLMs, diffusion models and procedural planning models suggest that world modeling and procedural planning are still two tasks which frontier models largely struggle with, despite humans easily solving both tasks. Current best performing models largely rely on textual descriptions to tackle both tasks, especially procedural planning, whereas humans are able to solve the task from the observations alone. Filling this gap is essential to provide models with a better understanding of our world at a higher-level and enable future AI systems to assist humans in a variety of tasks.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Bar, A., Zhou, G., Tran, D., Darrell, T., and LeCun, Y. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Ben-Shabat, Y., Yu, X., Saleh, F., Campbell, D., Rodriguez-Opazo, C., Li, H., and Gould, S. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 847–859, 2021.

Bruce, J., Dennis, M. D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Chang, C.-Y., Huang, D.-A., Xu, D., Adeli, E., Fei-Fei, L., and Niebles, J. C. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pp. 334–350. Springer, 2020.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pp. 1–23, 2022.

Garrido, Q., Assran, M., Ballas, N., Bardes, A., Najman, L., and LeCun, Y. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., and LeCun, Y. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Li, Y., Zhang, G., and Xu, C. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

Gupta, S., Wang, C., Wang, Y., Jaakkola, T., and Jegelka, S. In-context symmetries: Self-supervised learning through contextual world models. *Advances in Neural Information Processing Systems*, 37:104250–104280, 2024.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., and Corrado, G. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Islam, M. M., Nagarajan, T., Wang, H., Chu, F.-J., Kitani, K., Bertasius, G., and Yang, X. Propose, assess, search: Harnessing llms for goal-oriented planning in instructional videos. In *European Conference on Computer Vision*, pp. 436–452. Springer, 2024.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Kim, Y., Singh, G., Park, J., Gulcehre, C., and Ahn, S. Imagine the unseen world: a benchmark for systematic generalization in visual world models. *Advances in Neural Information Processing Systems*, 36:27880–27896, 2023.

Koh, J. Y., Lee, H., Yang, Y., Baldridge, J., and Anderson, P. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14738–14748, 2021.

LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Li, Z., Geng, W., Li, M., Chen, L., Tang, Y., Lu, J., and Zhou, J. Skip-plan: Procedure planning in instructional videos via condensed action space learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10297–10306, 2023.

Liu, J., Li, S., Wang, Z., Li, M., and Ji, H. A language-first approach for procedure planning. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1941–1954, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.122. URL https://aclanthology.org/2023.findings-acl.122/.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T., and Florence, P. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

Mendonca, R., Bahl, S., and Pathak, D. Structured world models from human videos. 2023.

Niu, Y., Guo, W., Chen, L., Lin, X., and Chang, S.-F. Schema: State changes matter for procedure planning in instructional videos. In *The Twelfth International Conference on Learning Representations*, 2024.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

Patel, D., Eghbalzadeh, H., Kamra, N., Iuzzolino, M. L., Jain, U., and Desai, R. Pretrained language models as visual planners for human assistance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15302–15314, 2023.

Shah, D., Osiński, B., Levine, S., et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pp. 492–504. PMLR, 2023.

Sun, J., Huang, D.-A., Lu, B., Liu, Y.-H., Zhou, B., and Garg, A. Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 7(2):4924–4930, 2022.

Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.

Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., and Zhou, J. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, A.-L., Lin, K.-Y., Du, J.-R., Meng, J., and Zheng, W.-S. Event-guided procedure planning from instructional videos with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13565–13575, 2023a.

Wang, H., Wu, Y., Guo, S., and Wang, L. Pdpp: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14836–14845, 2023b.

Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., and Lu, J. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024a.

Wang, Y., He, J., Fan, L., Li, H., Chen, Y., and Zhang, Z. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024b.

Wu, P., Escontrela, A., Hafner, D., Abbeel, P., and Goldberg, K. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023.

Wu, Y., Wang, H., Wang, J., and Wang, L. Open-event procedure planning in instructional videos. *arXiv preprint arXiv:2407.05119*, 2024.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), Online, November 2021. Association for Computational Linguistics.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. Qwen2 technical report. *CoRR*, 2024.

Yang, S., Du, Y., Ghasemipour, S. K. S., Tompson, J., Schuurmans, D., and Abbeel, P. Learning interactive real-world simulators. In *NeurIPS 2023 Workshop on Generalization in Planning*, 2023.

Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., and Zhou, J. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

Zhao, H., Hadji, I., Dvornik, N., Derpanis, K. G., Wildes, R. P., and Jepson, A. D. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2938–2948, 2022.

Zhou, G., Pan, H., LeCun, Y., and Pinto, L. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024a.

Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.-Y., and Gan, C. Robodreamer: learning compositional world models for robot imagination. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 61885–61896, 2024b.

Zhukov, D., Alayrac, J.-B., Cinbis, R. G., Fouhey, D., Laptev, I., and Sivic, J. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3537–3545, 2019.

A. Additional Information on Human Annotations

A.1. Human Annotation Statistics

In this section, we provide additional information concerning the human evaluation setup. A total of 34 annotators for World Modeling and 46 annotators for Procedural Planning were asked to solve the initial total of 1500 samples for each tasks, while ensuring that each sample will be solved by two different annotator. We ask that annotators should work on a minimum of 20 samples to have time to acclimate themselves to each task, and a maximum of 100 samples to avoid diminishing attention and quality. This resulted in each annotator solving an average of 88 samples for World Modeling, and 65 samples for Procedural Planning, which is effectively more difficult and time-consuming to solve. We provide the inter-annotator agreement on the original split of the benchmark for both tasks in table 2, with 73% on World Modeling and 65% on Procedural Planning, showing substantial agreement and reliability of the annotation results.

Dataset	# Annotators	Avg. # Samples per Annotator	Inter-Annotator Agreement
WorldPrediction-WM	34	88	0.73
WorldPrediction-PP	46	65	0.65

Table 2. Number of annotators, average number of samples evaluated per annotator and inter-annotator agreement for the human evaluation and filtering.

A.2. Human Annotation Setting

Before starting the annotation task, as the tasks can be conceptually confusing for humans due to the use of action equivalents, each annotator is given four solved examples of World Modeling and two solved examples of procedural planning along with the explanation of how to choose the correct candidate. One solved example for World Modeling is shown in Figure A.3 and a solved example for Procedural Planning is shown in Figure A.3. Along with the solved examples, the annotators are given the following in-depth instructions:

World Modeling Instruction for Human Annotation

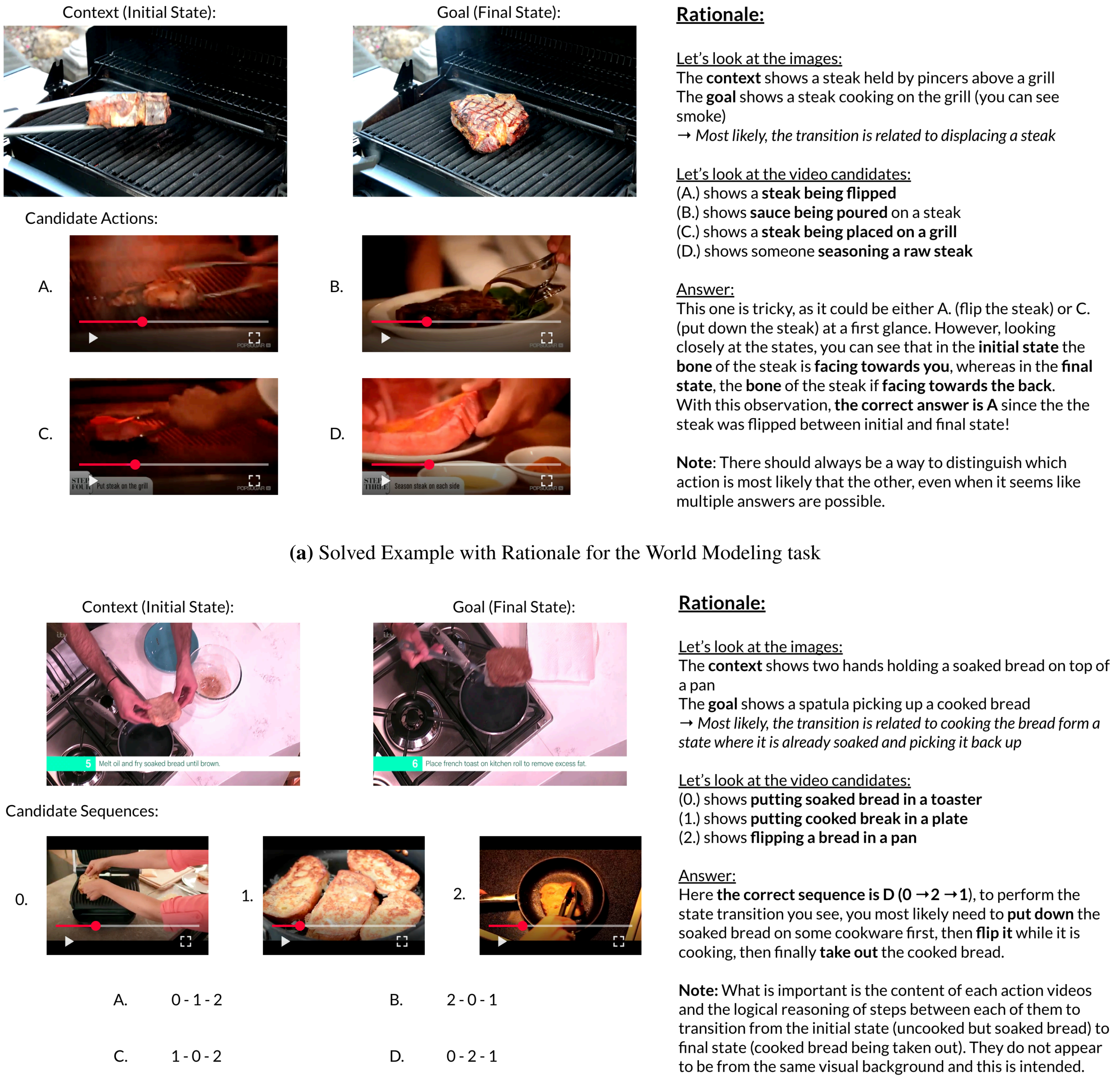
For the World Modeling task, you'll see two images showing a ***"before"****, as context, and an* ***"after"****, as goal, situation (for example, an empty cooking pot as "before", and a cooking pot containing water as "after").* ***Your job is to select which one of the four provided videos correctly shows the action performed to transition from the first initial state image to the second final state image.*** *Please pay attention to the action itself instead of the visual background (scenery or objects). We intentionally sampled the videos to depict the actions performed in a completely different environment (continuing the last example, the correct video answer could be showing a different liquid, like milk, being poured in a different pot: what matters is the performed action itself, here it would have been "Pouring liquid into container").*

Procedural Planning Instruction for Human Annotation

For the Procedural Planning task, you'll see two images showing a ***"before"****, as context, and an* ***"after"****, as goal, situation (for example, ingredients laid out separately, and then a finished sandwich).* ***Your job is to select which one of the provided sequences of videos (each consisting of several short video clips) correctly shows the correct order of action sequence to transition from the first initial state to the second final state image.*** *Please pay attention to the actions themselves instead of the visual background (scenery or objects), as we intentionally selected videos depicting the correct actions but performed in completely different environments (continuing the last example, the correct sequence could be something like (1) put the ham on some bread (2) put the cheese (3) close the sandwich, but each action could be depicted in a different environment)*

A.3. Human Annotation Solved Examples

We show here the solved examples that were presented to the human annotators to better understand the task. Each solved example was provided with a detailed explanation on how to solve the sample in question. Each annotator had 4 solved examples of WorldPrediction-WM and 2 solved examples of WorldPrediction-PP shown to them, according to which task they were annotating.



Rationale:

Let's look at the images:
The **context** shows a steak held by pincers above a grill
The **goal** shows a steak cooking on the grill (you can see smoke)
→ *Most likely, the transition is related to displacing a steak*

Let's look at the video candidates:
(A.) shows a **steak being flipped**
(B.) shows **sauce being poured** on a steak
(C.) shows a **steak being placed on a grill**
(D.) shows someone **seasoning a raw steak**

Answer:
This one is tricky, as it could be either A. (flip the steak) or C. (put down the steak) at a first glance. However, looking closely at the states, you can see that in the **initial state** the **bone** of the steak is **facing towards you**, whereas in the **final state**, the **bone** of the steak if **facing towards the back**.
With this observation, **the correct answer is A** since the the steak was flipped between initial and final state!

Note: There should always be a way to distinguish which action is most likely that the other, even when it seems like multiple answers are possible.

(a) Solved Example with Rationale for the World Modeling task

Rationale:

Let's look at the images:
The **context** shows two hands holding a soaked bread on top of a pan
The **goal** shows a spatula picking up a cooked bread
→ *Most likely, the transition is related to cooking the bread form a state where it is already soaked and picking it back up*

Let's look at the video candidates:
(0.) shows **putting soaked bread in a toaster**
(1.) shows **putting cooked break in a plate**
(2.) shows **flipping a bread in a pan**

Answer:
Here **the correct sequence is D (0 →2 →1)**, to perform the state transition you see, you most likely need to **put down** the soaked bread on some cookware first, then **flip it** while it is cooking, then finally **take out** the cooked bread.

Note: What is important is the content of each action videos and the logical reasoning of steps between each of them to transition from the initial state (uncooked but soaked bread) to final state (cooked bread being taken out). They do not appear to be from the same visual background and this is intended.

(b) Solved Example with Rationale for the Procedural Planning task

Figure 3. (a) and (b) illustrate the solved examples and rationales for WorldPrediction-WM and WorldPrediction-PP that were shown to annotators to guide their evaluations.

B. WorldPrediction Theoretical Formulation

B.1. Semi Partially-Observable Markov Decision Process Framework for WorldPrediction

We begin by formally defining a mathematical framework that provides the foundation for building the WorldPrediction benchmark. This formulation integrates elements from Partially Observable MDPs (Kaelbling et al., 1998) and Semi-MDPs (Sutton et al., 1999) to accurately capture the complex dynamics inherent in human activity videos. Formally, we represent this framework as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O} \rangle$:

World States $s \in \mathcal{S}$ constitute the continuous latent space representing the full underlying configuration of the environment. These states, although comprehensive, cannot be directly accessed and must instead be inferred from partial visual observations. Crucially, not all elements of a state are equally relevant to a given task: we distinguish between **task-relevant** components, which directly affect the causal outcomes of actions and are essential for achieving goals, and **task-irrelevant** components, representing background details or contextual information that do not influence the task.

(High-level) Actions $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$ represent the vocabulary of all possible actions. Here, "high-level" is characterized by both **semantic and temporal abstraction**, differentiating them from low-level continuous controls executed at fixed intervals. Each high-level action encapsulates several lower-level motor primitives or sub-actions. This can be modeled by *options* in Semi-MDPs, which is defined by a policy over low-level primitives, a termination condition, and a set consisting world states that allows that specific action. All components are dependent on the current environmental states, ensuring adaptation to varying contexts, as illustrated in Fig. 8. To distinguish from abstract action categories, we use the notation $a \in A$ to represent an action instance performed in a specific context s (*e.g.,* A_i represents "cut potato" and $a \in A_i$ is the muscle motion sequence of cutting potato in one particular kitchen settings).

Transition Model $\mathcal{T}$ specifies the true underlying mechanism governing how world states evolve over time – after an action a_t being taken at s_t, the world state transit to a new state s_{t+1} with a probability of $\mathcal{T}\big(s_{t+1} \mid s_t, a_t\big)$. In real-world, non-simulated environments, this transition mechanism is hidden and thus inaccessible; agents must approximate it by learning a **world model**. It enables reasoning and planning without relying directly on explicit reward signals or costly trial-and-error interactions in the real world.

Observation Model $\mathcal{O}$ maps latent world states or performed actions to corresponding sensory signals, *i.e.,* an image $\mathcal{O}(s_t)$ and a video segment $\mathcal{O}(a_t)$. Due to intrinsic limitations of perception devices (*e.g.,* occlusions, resolution, or viewpoint constraints), they only provide imperfect views of the underlying true state or the performed action, and also contain excessive amount of task-irrelevant background information. To address these challenges brought by **partial observability**, our benchmark incorporates two strategies detailed in §B.3: *observability filtering*, which excludes samples lacking sufficient visual evidence of action outcomes, and *action equivalents*, which mitigate the shortcut based on superficial background continuity cues.

Given the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O} \rangle$, we can formally characterize the underlying data generative process of human activity videos as follows. Beginning from an initial latent state s_0, a human agent decided to perform an action $a_0 \in A_i$. The transition model $\mathcal{T}$ subsequently generates the next latent state s_1 conditioned on s_0 and a_0. This process iterates over multiple steps. Through the observation model $\mathcal{O}$, each latent state s_t and action a_t is mapped to visual observations, yielding the observed video sequence: $[\mathcal{O}(s_0), \mathcal{O}(a_0), \mathcal{O}(s_1), \mathcal{O}(a_1), \ldots, \mathcal{O}(s_T)]$.

B.2. Benchmark Objectives

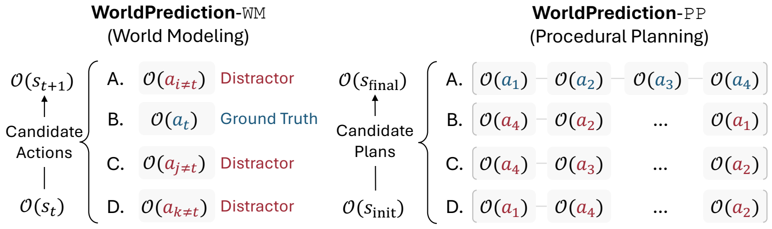

WorldPrediction-WM
(World Modeling)
$\mathcal{O}(s_{t+1})$
Candidate Actions
$\mathcal{O}(s_t)$
A. $\mathcal{O}(a_{i\neq t})$ Distractor
B. $\mathcal{O}(a_t)$ Ground Truth
C. $\mathcal{O}(a_{j\neq t})$ Distractor
D. $\mathcal{O}(a_{k\neq t})$ Distractor
WorldPrediction-PP
(Procedural Planning)
$\mathcal{O}(s_{\text{final}})$
Candidate Plans
$\mathcal{O}(s_{\text{init}})$
A. $\mathcal{O}(a_1)$ $\mathcal{O}(a_2)$ $\mathcal{O}(a_3)$ $\mathcal{O}(a_4)$
B. $\mathcal{O}(a_4)$ $\mathcal{O}(a_2)$ ... $\mathcal{O}(a_1)$
C. $\mathcal{O}(a_4)$ $\mathcal{O}(a_3)$ ... $\mathcal{O}(a_2)$
D. $\mathcal{O}(a_1)$ $\mathcal{O}(a_4)$ ... $\mathcal{O}(a_2)$

Figure 4. **Discriminative task formulation of WorldPrediction**. Each sample includes a pair of visual observation of states along with a set of candidate actions or action sequences. Models must identify the correct one responsible for the observed state transition among distractors. Note that every $\mathcal{O}(a)$ is substituted by its action equivalent to avoid trivial background continuity shortcut.

Our primary goal is to measure a model's ability to understand real-world state transitions and the causal factors that drive them. Concretely, we focus on capturing how an initial world configuration evolves to a new configuration when subjected to a particular high-level action. This predictive ability, known as *world modeling*, is formalized by having a learned function $\mathcal{W}$ to approximate the true underlying transition model $\mathcal{T}$. Under a suitable divergence metric $\mathcal{D}$, the performance of a world model can be naturally defined as:

$$\mathcal{D}\left(\mathcal{W}(s_{t+1} \mid s_t, a_t) \,\|\, \mathcal{T}(s_{t+1} \mid s_t, a_t)\right). \tag{1}$$

Intuitively, a high-performing world model assigns higher likelihood to correct state transitions $(s_t, a_t) \to s_{t+1}$ and lower likelihood to incorrect transitions involving counterfactual combinations of states and actions. Formally, given a learned transition model $\mathcal{W}$, this implies the inequality $\mathcal{W}(s_{t+1} \mid s_t, a_t) > \mathcal{W}(s_{t+1} \mid s_t, a_j)$ for any counterfactual action $a_j \neq a_t$. Because we specifically focus on evaluating understanding of *high-level* actions rather than low-level primitives, we define this criterion at the action-category level: given the true action category A^* corresponding to the correct action a_t, we empirically approximate the theoretical divergence by verifying whether the model assigns the highest likelihood to the correct action category responsible for the observed transition:

$$A^* \overset{?}{=} \arg\max_{A \in \mathcal{A}} \mathcal{W}(s_{t+1} \mid s_t, A). \tag{2}$$

This formulation probes a model's approximation of the hidden transition model $\mathcal{T}$ by evaluating how well the causal relationship between (s_t, a) and s_{t+1} is captured. To have a robust approximation of $\mathcal{T}$, world models should learn to capture and discriminate the various ways in which actions transform the latent world state, rather than simply matching superficial or spurious correlations between states and actions, which we ensure in our design detailed later in section B.3.

This argmax formulation of evaluation also enables a natural extension to multi-step procedural planning evaluation, where a *plan* consisting sequence of actions can be viewed as a single *"macro-action"*, linking distant initial and final states. Specifically, given an initial state s_{init} and a final state s_{final} separated by T high-level actions, the objective is to select the correct *ordered* sequence of actions $\mathcal{P}^* = (a_1, \ldots, a_T)$ responsible for this long-horizon transition:

$$\mathcal{P}^* \overset{?}{=} \arg\max_{\mathcal{P} \in \mathcal{A}^T} \mathcal{W}\big(s_{\text{final}} \mid s_{\text{init}}, \mathcal{P}\big), \tag{3}$$

where $\hat{\mathcal{P}} = (\hat{a}_1, \ldots, \hat{a}_T)$ denotes the correct action sequence that transit s_{init} to s_{final}, and $\mathcal{A}^T$ denotes candidate plans of all possible arrangement of T-step action sequence. In principle, if all intermediate states $(s_2, \ldots, s_{T-1})$ were known, solving procedural planning would reduce to solving T successive world modeling steps. However, since these intermediate states are unobserved, the model must internally infer them, effectively reasoning about the entire multi-step causal chain.

B.3. Benchmark Design

Task Formulation. We now outline the design of our benchmark. As the true underlying states and transitions in real-world scenarios are not directly accessible, our benchmark instead leverages *visual observations*— images or video clips—as cues to infer the true states and actions. We present WorldPrediction-`WM` and WorldPrediction-`PP`, two benchmarks respectively evaluating world modeling (Eq. 2) and procedural planning (Eq. 3) capabilities. Concretely, each sample consist of:

- **State Observations**: Static images capturing the environment's configuration before and after the action(s) being taken, denoted as $\mathcal{O}(s_t)$, $\mathcal{O}(s_{t+1})$ for WorldPrediction-`WM` and $\mathcal{O}(s_{\text{init}})$, $\mathcal{O}(s_{\text{final}})$ for WorldPrediction-`PP`.
- **Action / Plan Candidates**: The search space of the argmax operation in Eq. 2 and Eq. 3, containing one ground truth (A^* or $\mathcal{P}^*$) and many distractors. To enhance computational efficiency, the candidate pool can be limited to a small subset of the complete action space $\mathcal{A}$ or plan space $\mathcal{A}^T$.

Models must select which action (or action sequence) accounts for the observed change in $\mathcal{O}(s_t) \to \mathcal{O}(s_{t+1})$ or $\mathcal{O}(s_{\text{init}}) \to \mathcal{O}(s_{\text{final}})$ providing a clear evaluation of world modeling and procedural planning. This discriminative multiple-choice (illustrated in Fig. 4) setup directly aligns with our theoretical grounding (Eq. 2 and Eq. 3), and also offers several practical advantages. It universally accommodates different types world models and planners (*e.g.,* models using different architectures, generating different modalities to represent the predicted states). Additionally, by using only raw visual observations, we remove the reliance on human annotated text labels as done in previous benchmarks (Chang et al., 2020), ensuring an unbiased evaluation[1].

Action Equivalents. Due to being purely observation-based, an important challenge in the construction of our benchmark is to prevent models from exploiting trivial continuity cues to identify the correct action or sequence. Specifically, if the same camera viewpoint, background objects, or other task-irrelevant visual elements are preserved across the state observations as well as the ground-truth action segment, then a model might simply match low-level features without learning the true causal relationship between action content and state transitions. Such an approach would results in models failing to capture the *semantic and temporal abstractions* of high-level actions. To mitigate this shortcut, we employ **action equivalents** (shown in Appendix, Fig. 8). For each high-level action category A_i, there exits a set of observations which depict it being performed in visually different environments or from a significantly different viewpoint (*e.g.,* egocentric vs exocentric). Concretely, we use that set to replace the ground-truth observation action with one of its action equivalent and re-sample distractors from the same environment of the action equivalent for WorldPrediction-`WM`, and re-shuffle the new sequence of equivalent actions for WorldPrediction-`PP`.

Observability Filtering. Under the *partial observability* assumption, task-relevant elements of the environment can sometimes fail to be captured in state observations. When the evidence needed to infer what changed—and thus which action caused the transition— is missing, the ambiguity increases significantly and the task becomes nearly impossible even for humans. There are two main causes for failing to capture the action-relevant state observation: **noisy observation** due to video edits or drastic camera field-of-view shifts, and **occlusions** due to different entities blocking the view of task-relevant objects. In this section, we present our solutions for filtering out those low-quality samples.

To remove samples with noisy observation, we employ an assumption that the noisy observation usually causes larger changes in semantic feature space. Specifically, we compute the distance d between the visual features for both state observations $d = \|\phi(\mathcal{O}(s_{\text{init}})) - \phi(\mathcal{O}(s_{\text{final}}))\|_2$ using a pretrained vision encoder $\phi(\cdot)$ and we only keep pairs $(\mathcal{O}(s_{\text{init}}), \mathcal{O}(s_{\text{final}}))$ whose similarity score is smaller than certain threshold, thus removing samples where the scene changes so drastically that no coherent causal link can be reliably inferred. The left side in Fig. 7 provide example of this filtering.

This filtering process can be seen as a coarse classifier that eliminates a large portion of the bad state observations by relying on the assumption that observations which are too different are highly likely to miss task-relevant information in at least one of the two states. This assumption also aligns with the POMDP formulation: consecutive observations of the same environment should not appear uncorrelated if they reflect smoothly evolving states in the real-world.

Additionally, we filter out exocentric state observation where the human performing the action has his back turned toward the camera (or otherwise heavily obstructing the view, as shown in the bottom-right of Fig. 7), as in such cases it becomes exceedingly difficult to discern the critical objects or interactions relevant to the action. Consequently, the remaining samples more consistently capture the essential task-relevant cues for modeling and evaluating high-level transitions, aligning with the **partial observability** principle in a controlled yet realistic setting.

[1] Although models can still generate captions from visual observations (as in Socratic LLM baselines provided in §4), we view them as models' internal perceptual representations.

C. Additional Dataset Information

Dataset	WorldPrediction-WM			WorldPrediction-PP		
	# Samples	# Unique Actions	Avg. Duration (s)	# Samples	# Unique Actions	Avg. Duration (s)
COIN	236	532	13.16	243	285	14.70
CrossTask	109	194	9.17	58	65	7.53
IKEA ASM	159	185	9.02	136	43	6.48
EgoExo4D	128	128	11.71	76	180	11.23
EPIC-KITCHENS-100	193	561	6.25	57	176	3.47
WorldPrediction (All)	825	1800	10.02	570	749	9.38

Table 3. WorldPrediction dataset statistics (number of samples, actions, and average action duration) for both tasks

Dataset Sources. WorldPrediction incorporates five publicly available datasets to ensure broad coverage and representativity of skilled human activities:

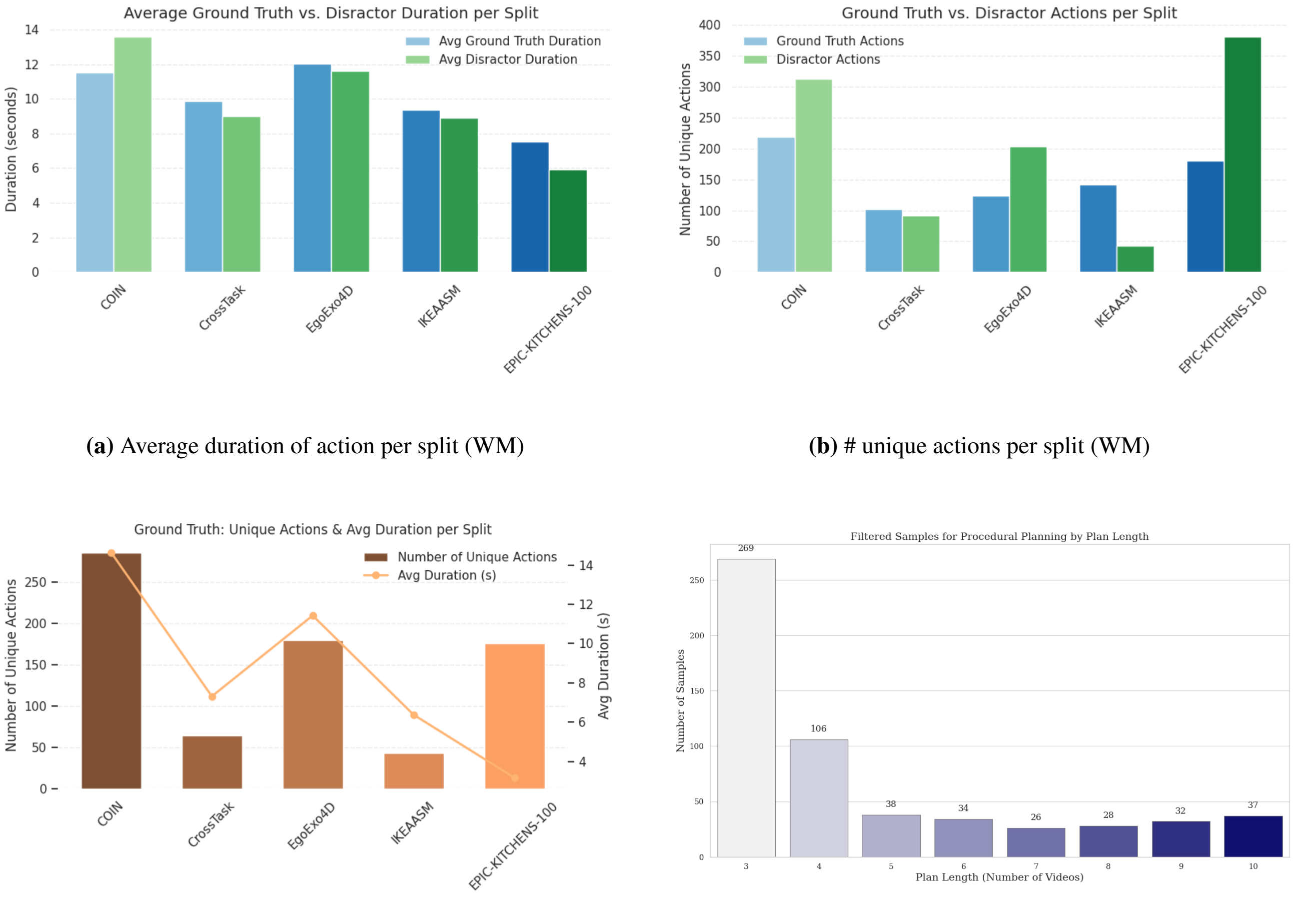
- **COIN** (Tang et al., 2019): provides instructional web videos covering diverse procedural tasks, such as cooking and household repairs.
- **CrossTask** (Zhukov et al., 2019) consists of instructional web videos capturing diverse everyday activities.
- **EgoExo4D** (Grauman et al., 2024) provides temporally-aligned egocentric and multi-view exocentric videos. We focus specifically on the `cooking` and `healthcare` subsets, which emphasize procedural human activities.
- **EPIC-KITCHENS-100** (Damen et al., 2022): is a large-scale egocentric dataset of kitchen tasks with detailed annotations, capturing fine-grained interactions.
- **IKEA-ASM** (Ben-Shabat et al., 2021) features clear exocentric instructional videos of furniture assembly, providing structured action sequences in controlled environments.

For the WorldPrediction-WM task, we show the average duration of the ground truth action vs the average duration of the distractor actions per dataset split in Figure 5.a and the number of unique actions that appear as ground truth and as distractors per dataset split in Figure 5.b. Similarly for the WorldPrediction-PP task, as the distractors are shuffled version of the same actions, we directly show the unique actions and average duration per dataset split in Figure 5.c.

EPIC-KITCHENS-100 have relatively shorter action observations for both World Modeling and Procedural Planning, this is expected as the original dataset contain a limited amount of samples but extremely fine-grained annotation of actions (e.g., *pick up, put down, open*) while actions in dataset like COIN and EgoExo4D are more macroscopic (e.g. *add, mix, boil*). This is also interesting for obtaining more robust results on our benchmark, as the duration of the action clips is not standardized and hence does not favor any types of models.

The number of unique action in IKEAASM and CrossTask is smaller than other datasets for two reasons: first because the number of samples are smaller as shown in 3 due to the human filtering, but also because for IKEAASM for example, the action space is very limited as the dataset only contains four different types of furnitures, so the action overlap is significant. This is not a problem in our benchmark as the assembly domain is proportionally well represented, and some of the CrossTask domains overlap with COIN's domains. Finally, we show the number of samples per plan length in Figure 5.d, with a majority of plans of length 3 and 4 to reflect current planning datasets, but with a uniform number of samples for plans from 5 to 10 with a bit more than 30 on average.

We also provide a visualization of the 50 most frequent actions appearing in both the World Modeling and the Procedural Planning tasks in Figure 6. As the original filtering to deem a World Modeling sample valid vs. a Procedural Planning sample valid differs, the distribution for the action frequency is also different. The action annotations are also provided in the benchmark dataset for researchers interested in only specific domains, tasks or actions. Due to the very small action space of IKEAASM, we choose not to display the actions belonging to the aforementioned split for the figure to be easier to read. The action information concerning IKEAASM can be found on the released dataset benchmark.



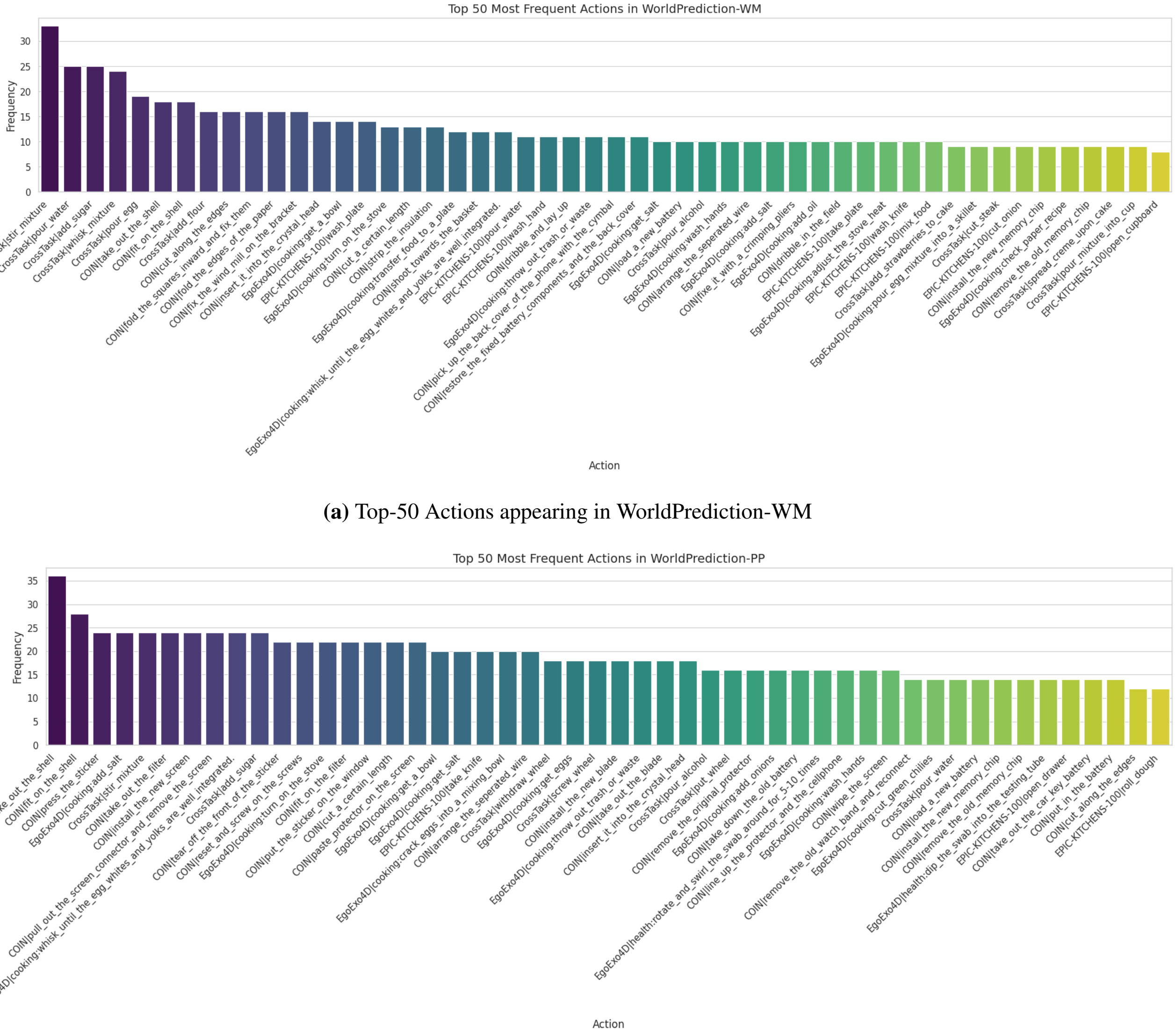
(a) Average duration of action per split (WM)

(b) # unique actions per split (WM)

(c) Average duration and # actions per split (PP)

(d) # Samples per Plan Length (PP)

Figure 5. Additional dataset information: average duration of actions and number of actions per split for both tasks, and number of samples per plan length in Procedural Planning.



(a) Top-50 Actions appearing in WorldPrediction-WM

(b) Top-50 Actions appearing in WorldPrediction-PP

Figure 6. Top-50 most frequent actions across WorldPrediction-WM and WorldPrediction-PP datasets (excluding IKEA ASM due to the small action space yielding very high frequency of assembly actions).

D. Additional information on Sample Filtering & Action Equivalents

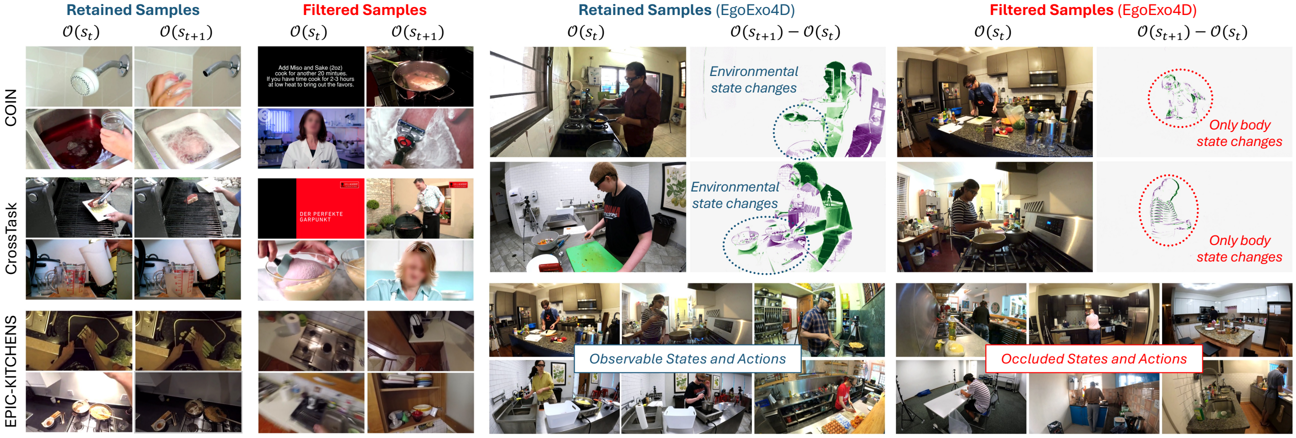

Retained Samples
$\mathcal{O}(s_t)$
$\mathcal{O}(s_{t+1})$
Filtered Samples
$\mathcal{O}(s_t)$
$\mathcal{O}(s_{t+1})$
Retained Samples (EgoExo4D)
$\mathcal{O}(s_t)$
$\mathcal{O}(s_{t+1}) - \mathcal{O}(s_t)$
Filtered Samples (EgoExo4D)
$\mathcal{O}(s_t)$
$\mathcal{O}(s_{t+1}) - \mathcal{O}(s_t)$
COIN
CrossTask
EPIC-KITCHENS
Add Miso and Sake (2oz) cook for another 20 mintues. If you have time cook for 2-3 hours at low heat to bring out the favors.
DER PERFEKTE GARPUNKT
Environmental state changes
Environmental state changes
Only body state changes
Only body state changes
Observable States and Actions
Occluded States and Actions

Figure 7. **Sample Filtering in WorldPrediction.** Samples are retained only if state observations clearly show meaningful environmental changes resulting from actions. Samples are filtered out if they exhibit excessive viewpoint shifts, only contain minor body movements without clear environmental changes, or severe occlusions, which all makes causal inference challenging.

For samples with different camera angles (mostly present in EgoExo4D), we further filter samples in which the human subject is obstructing the task-relevant objects, in which case it is impossible to either see the action or the consequences of the action on the object states. We further these samples by simply prompting a VLM (Qwen2.5-VL 72b) with: `"Is the main person not showing their back and what they are doing with hands being clearly visible?"`. Note that this is done prior to the human validation round to discard completely impossible samples and reduce human annotator's workload, but the human annotation round would have very likely filtered these samples out anyways.

We show some action equivalents in Fig. 8 and the generation for a sample of WorldPrediction-`WM` using CogVideoX-I2V in Fig. 9. Despite minor artifacts, the generations themselves make sense. As for generation models, hallucination still happens quite often. For ”Add Ice” for example, the model seems to understand some form of liquid –most probably alcohol, although square ice is probably what was meant in the original label. The final mixture color is also not entirely correct, but these artifacts are not the main source of errors for diffusion models. The current evaluation principle naively compares the last generated frame with the final state observation, which is not assured by any hyperparameters.

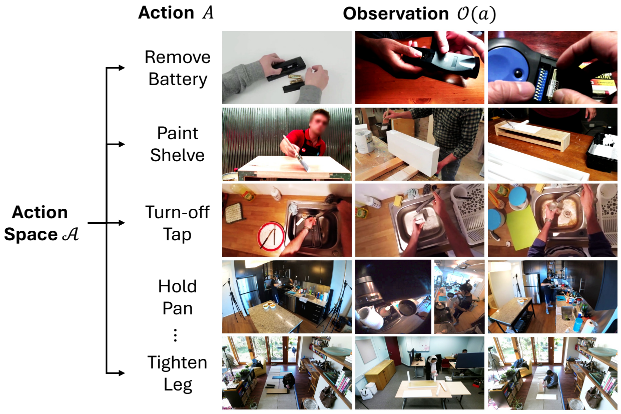


Figure 8. **High-level Actions in WorldPrediction.** The action space $\mathcal{A}$ consists of abstract action categories A, each instantiated through multiple specific actions a performed across different environments. Each action is represented as a video clip $\mathcal{O}(a)$ (The textual labels are for illustration purposes only and are not included in the benchmark).

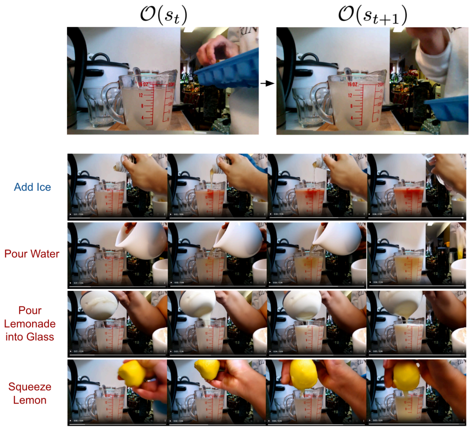


Figure 9. **Video generated by diffusion world model.** Example of the generation from CogVideoX-I2V of the four action prompt given $\mathcal{O}(s_t)$.

E. Additional Information on Models & Evaluations

E.1. Models

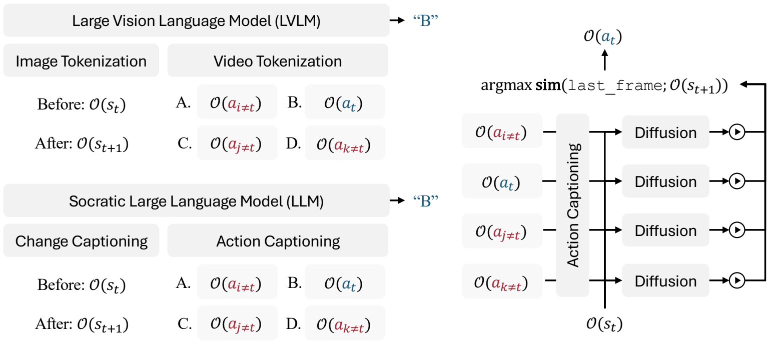

Large Vision Language Model (LVLM)
“B”
Image Tokenization
Video Tokenization
Before: $\mathcal{O}(s_t)$
A. $\mathcal{O}(a_{i\neq t})$
B. $\mathcal{O}(a_t)$
After: $\mathcal{O}(s_{t+1})$
C. $\mathcal{O}(a_{j\neq t})$
D. $\mathcal{O}(a_{k\neq t})$
Socratic Large Language Model (LLM)
“B”
Change Captioning
Action Captioning
Before: $\mathcal{O}(s_t)$
A. $\mathcal{O}(a_{i\neq t})$
B. $\mathcal{O}(a_t)$
After: $\mathcal{O}(s_{t+1})$
C. $\mathcal{O}(a_{j\neq t})$
D. $\mathcal{O}(a_{k\neq t})$
$\mathcal{O}(a_t)$
argmax **sim**(last_frame; $\mathcal{O}(s_{t+1})$)
$\mathcal{O}(a_{i\neq t})$
$\mathcal{O}(a_t)$
$\mathcal{O}(a_{j\neq t})$
$\mathcal{O}(a_{k\neq t})$
Action Captioning
Diffusion
Diffusion
Diffusion
Diffusion
$\mathcal{O}(s_t)$

Figure 10. **Baseline models**. VLMs directly encode visual observations, while Socratic LLMs first generate textual captions describing state changes and candidate actions, then select the action through text-only reasoning. Video diffusion models generate future observations conditioned on action captions, selecting the action by comparing final generated frame and the desired $\mathcal{O}(s_{t+1})$.

VLMs. We use two state-of-the-art open-source VLMs families: `Qwen2.5-VL` (Yang et al., 2024) and `InternVL2.5` (Chen et al., 2024). As shown in Fig. 10, to perform the WorldPrediction multi-choice task, models are prompted with a structured multimodal query comprising images depicting the initial and final world states, video segments representing the candidate actions, along with textual instruction explaining the task and specifying the desired output format. We frame the task explicitly by instructing the model to select the most plausible action or the sequence of actions that cause the observed state transition.

Socratic LLMs. We evaluate the performance of Socratic LLMs (Zeng et al., 2022), which decouple perception and reasoning into two distinct stages. Visual inputs are translated into textual descriptions through a VLM, then a text-only instruct-tuned LLM is prompted with these captions along with instructions, including structured task explanations and candidates. The LLM then employs textual reasoning to identify the action or sequence of actions most plausibly causing the observed state transitions. To obtained the textual description, we utilized Qwen 2.5-VL (72B). For text-only LLM, we evaluated five different LLM families with varying sizes, including Llama 3.1-Instruct (8B, 70B, 405b), Qwen 2.5-Instruct (3B, 7B, 14B, 72B), DeepSeekR1 (distilled version Qwen-32B), GPT-4o and Claude-3.5-Sonnet.

Video Diffusion Models. To assess generative world modeling capabilities, we also evaluate two image-conditioned video diffusion models: I2VGenXL (Zhang et al., 2023) and CogVideoX-I2V (Hong et al., 2022) which directly generates the future state in pixel space. For inference, we provide the initial state observation $\mathcal{O}(s_t)$ as the grounding image, and perform action captioning using a VLM to get a text description of each action candidates. The generated video is a visual representation of the state transition towards the final state observation $\mathcal{O}(s_{t+1})$. We select the most likely action candidate by identifying the generated segment whose last frame exhibits the smallest pixel-wise distance to $\mathcal{O}(s_{t+1})$.

OEPP Models. We reimplement OEPP models (Wu et al., 2024) and incorporate them into the WORLDPREDICTION-PP task. OEPP performs planning using VideoCLIP (Xu et al., 2021) embeddings. Given initial and final observations, a planning model (either MLP, Transformer (Vaswani et al., 2017), or PDPP (Wang et al., 2023b)) is trained to generate T text embeddings corresponding to a sequence of T predicted actions. We embed all candidate plans into the same text embedding space and select the candidate that minimizes the distance with the generated embeddings.

E.2. Ablation on Captions for Socractic LLMs

Captioner	InternVL2.5			Qwen2.5VL			Oracle
	4B	8B	26B	3B	7B	72B	Captions
WorldPrediction-WM	33.3	39.0	46.3	35.2	40.3	48.5	56.0

Table 4. Comparing different captioners for Socratic LLM (Qwen2.5-72B-Instruct). Larger VLMs enable higher accuracy (%) on WorldPrediction-WM, but still a major gap remains to human-annotated action labels (oracle captions).

To further investigate the impact of caption quality on Socratic LLM performance, we conducted an ablation study summarized in Table 4. Specifically, we varied the VLM model used for generating textual descriptions of states and actions. As the captioning VLM model size increases, we observe a clear improvement in Socratic LLM performance, confirming that richer and more accurate textual captions significantly facilitate better textual reasoning. However, there is still a significant gap between best captioner Qwen2.5-VL (72B) and oracle captions (human-annotated labels).