Test-Time Adaptation for Generalizable Task Progress Estimation

Christos Ziakas [1] **Alessandra Russo** [1]

Abstract

We propose a test-time adaptation method that enables a progress estimation model to adapt online to the visual and temporal context of test trajectories by optimizing a learned self-supervised objective. We introduce a gradient-based meta-learning strategy that explicitly trains the progress estimator to learn to adapt at test time, inducing reliance on semantic content rather than temporal order. Our method generalizes from a single training environment to diverse out-of-distribution tasks, environments, and embodiments, outperforming the state-of-the-art in-context learning approach using autoregressive vision-language models.

1. Introduction

Vision-Language Models (VLMs) have demonstrated remarkable generalization capabilities across diverse tasks and domains by learning from large-scale, unstructured web data without human supervision (Radford et al., 2021; Alayrac et al., 2022). In contrast, despite significant advances in learning generalist policies for robotics (Brohan et al., 2023; Ghosh et al., 2024) and 3D virtual environments (Reed et al., 2022; Team et al., 2024), state-of-the-art methods in these domains have yet to achieve comparable success due to their reliance on expert demonstrations, limiting their scalability. World models have emerged as a promising solution (Ha & Schmidhuber, 2018; Hafner et al., 2021), enabling agents to learn in simulated environments. Despite the potential of world models to generate realistic visual trajectories in diverse, open-ended environments (Hughes et al., 2024; Silver & Sutton, 2025), a critical challenge remains: *how can agents effectively learn from videos at scale without relying on human supervision?* One promising approach is to learn a generalizable goal-conditioned value function directly from expert visual trajectories and their corresponding task descriptions, using it as a reward and supervision signal for reinforcement learning and imitation learning, respectively (Chen et al., 2021; Ma et al., 2022). In this framework, goal-conditioned value estimation has been formulated as *task progress estimation*: predicting how far an agent has progressed toward completing a task, based on visual observations and a natural language task description (Dashora et al., 2025).

Prior work on progress estimation has employed contrastive VLMs (Radford et al., 2021), leveraging the similarity between a task description and a visual observation in their shared multimodal representation. However, these methods do not incorporate temporal context from the visual trajectory during inference (Ma et al., 2023; Baumli et al., 2023; Rocamonde et al., 2024). In contrast, autoregressive VLMs (Alayrac et al., 2022) incorporate temporal context by conditioning on the full visual trajectory within the prompt; however, the state-of-the-art approach (Ma et al., 2024) discards temporal order by shuffling the trajectory in inference to mitigate a bias toward monotonically increasing predictions, inherited from chronologically ordered datasets used to pre-train VLMs. Both existing VLM architectures rely on pre-trained representations and in-context learning for generalization, which limits their robustness in tasks that require temporal reasoning (Pătrăucean et al., 2023).

In this work, we propose a method for training a progress estimation network that adapts to both visual and temporal contexts in unseen environments, using sequences of multimodal representations derived from expert visual trajectories and their corresponding task descriptions. At test time, instead of relying on similarity-based approaches or in-context learning, our method updates the parameters of the progress estimation network to adapt to context before making task progress predictions. As the network sequentially updates its parameters during inference, it captures temporal context, forming an implicit memory of the past (Sun et al., 2024). To mitigate reliance on temporal cues in progress estimation (Xu et al., 2024; Singh et al., 2021), we propose a gradient-based meta-learning strategy (Finn et al., 2017; Sun et al., 2020) that trains our model for test-time adaptation by optimizing the self-supervised loss over diverse sub-trajectories selected via dissimilarity-based sampling.

This work is supported by UKRI (EP/Y037111/1) as part of the ProSafe project (EU Horizon 2020, MSCA, grant no. 101119358). [1]Department of Computing, Imperial College London, London, United Kingdom. Correspondence to: Christos Ziakas <c.ziakas24@imperial.ac.uk>.

Accepted to the 2nd Workshop on Test-Time Adaptation: Putting Updates to the Test (PUT) at the 42^{nd} International Conference on Machine Learning (ICML), Vancouver, Canada, 2025.

2. Methodology

2.1. Problem Formulation

In video trajectories, task progress estimation is equivalent to learning a goal-conditioned value function under reward functions that measure task completion. Therefore, we formulate task progress estimation as the problem of learning a goal-conditioned value function that predicts the degree of task completion from visual-language representations. Formally, the value function is defined as: $V : \mathcal{O} \times \mathcal{G} \to [0, 1]$ which maps an observation $o_t \in \mathcal{O}$ and a goal specification $g \in \mathcal{G}$ to a scalar value indicating the predicted progress toward goal completion, where $V(o_t; g) = 0$ corresponds to the start of the task and $V(o_t; g) = 1$ to task completion. Task progress is often aligned with temporal position in expert demonstrations, based on the assumption that such trajectories exhibit monotonically increasing progress toward goal completion (Lee et al., 2021; Ma et al., 2024; Dashora et al., 2025). Given an expert trajectory $\tau = (o_1, \ldots, o_T) \sim \pi_E$, temporal progress is defined using normalized timestep indices: $V^{\pi_E}(o_t; g) = \frac{t}{T}$ where T is the trajectory length. This provides supervision for learning a goal-conditioned value function V, which can generalize beyond expert data to estimate task progress.

2.2. Model Architecture

Our method comprises three modules: (1) a multimodal encoder that extracts visual-language representations from visual trajectories and task descriptions; (2) a test-time adaptation module updated with a learned self-supervised objective; and (3) a regression head that predicts task completion.

2.2.1. Multimodal Input Representation

We use a frozen contrastive vision-language encoder, CLIP (Radford et al., 2021), to extract representations from visual observations and goal descriptions. Given a visual trajectory $\tau = (o_1, \ldots, o_T)$ and a language task description g, we concatenate their representations at each timestep t to form a sequence of joint multimodal representations $(x_1, x_2, \ldots, x_T)$, where $x_t = [\phi_v(o_t); \phi_g(g)] \in \mathbb{R}^{2d}$ and ϕ_v and ϕ_g denote the visual and language encoders, respectively. CLIP is pre-trained with a contrastive objective to align paired image-text inputs in a shared representation space. As a result, representations of visual observations that are semantically closer to goal completion tend to align with the representation of the goal description.

2.2.2. Test-Time Adaptation

To adapt the multimodal input representation to both visual and temporal context, we employ an adaptation module f_{adapt} following the test-time training paradigm (Sun et al., 2020). At inference time, the parameters of f_{adapt} are updated via a step of gradient descent on the self-supervised loss ℓ_{self}, based on the test trajectory. The self-supervised loss ℓ_{self} is parameterized by projection matrices, meta-learned by optimizing progress estimation performance after test-time adaptation. In particular, f_{adapt} is trained to reconstruct a projected target representation from a corrupted input generated by the projection matrices $P_V \in \mathbb{R}^{d' \times d}$ and $P_K \in \mathbb{R}^{d' \times d}$, respectively. At each timestep t, the model receives a sliding window of size $k + 1$ with the local representations $\mathcal{W}_{ctx} = \{x_{t-k}, \ldots, x_t\}$, and performs t_{ep} steps of gradient descent on ℓ_{self}:

$$\theta_t = \theta_{t-1} - \eta \sum_{x_\tau \in \mathcal{W}_{ctx}} \nabla_\theta \ell_{\text{self}}(x_\tau; \theta_{t-1}) \tag{1}$$

where η is the adaptation learning rate, and θ_{t-1} are the parameters from the previous step. For each trajectory, the parameters θ_0 learned during meta-learning serve as the initialization. The visual context is captured from the CLIP-based input representation x_t, while temporal context is captured through test-time adaptation over recent representations $x_{t-1}, x_{t-2}, \ldots$, either explicitly or implicitly (Wang et al., 2023). In the explicit case, the parameters of f_{adapt} are reset to θ_0, while in the implicit case they are incrementally updated from θ_{t-1} to retain information over time.

2.2.3. Task Progress Estimator

After test-time adaptation, a third projection matrix $P_Q \in \mathbb{R}^{d' \times d}$ maps the input x_t into an adaptation space $\mathbb{R}^{d'}$. Then, it is passed through an adaptation function f_{adapt}, followed by a progression head h to estimate task progress:

$$V(x_t; g) = h(f_{\text{adapt}}(P_Q x_t; \theta_t)) \tag{2}$$

where V is the predicted degree of task completion. The progression head h is a multilayer perceptron (MLP), following value function architectures used in deep reinforcement learning (Espeholt et al., 2018). The network is trained using normalized progress labels y_t from expert demonstrations, minimizing the mean squared error. In inference, the progression head h remains frozen; only the adaptation module f_{adapt} is updated using the self-supervised objective.

2.3. Training Procedure

The projection matrices P_K, P_V, P_Q, which parametrize the self-supervised loss ℓ_{self}, along with the progression head h are trained with gradient-based meta-learning to optimize the progress prediction objective $\mathcal{L}_{\text{pred}}$ after test-time adaptation (Sun et al., 2024). The total training objective combines $\mathcal{L}_{\text{pred}}$ with ℓ_{self}, weighted by a scalar λ, and is optimized by differentiating through the test-time adaptation steps applied to f_{adapt}. Progress estimators often exploit global temporal shortcuts (e.g., trajectory length) as a form of temporal shortcut bias (Xu et al., 2024; Singh et al., 2021), leading to poor

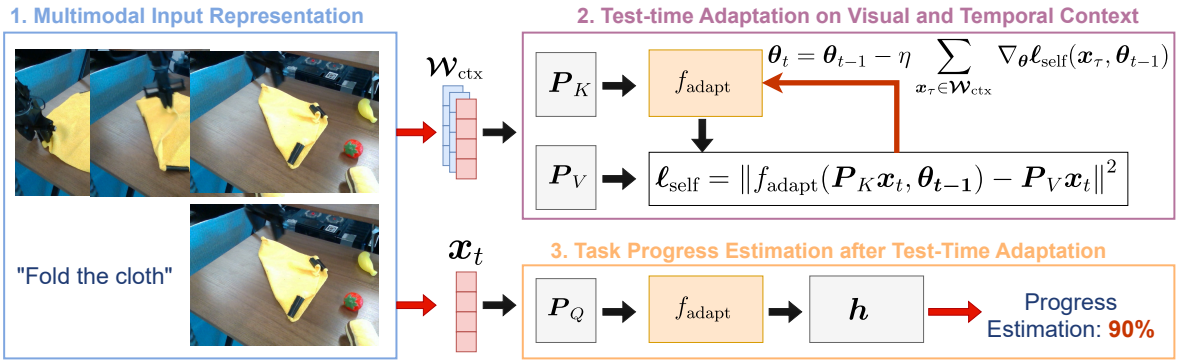


Figure 1: **Test-Time Adaptation for Progress Estimation.** 1) At each timestep, multimodal representation $\boldsymbol{x}_t$, and its context window $\mathcal{W}_{\text{ctx}} = \{\boldsymbol{x}_{t-k}, \ldots, \boldsymbol{x}_t\}$, serve as input to adaptation module f_{adapt}. 2) f_{adapt} is updated using self-supervised loss over $\mathcal{W}_{\text{ctx}}$. 3) The adapted representation is passed through frozen MLP head h to produce scalar task progress score.

performance. To encourage reliance on semantic rather than temporal cues, we apply test-time adaptation during training over diverse sub-trajectories selected via dissimilarity-based sampling. Given a sequence of multimodal representations $\{x_1, \ldots, x_T\}$, fixed-length sub-trajectories are extracted using a sliding window of size w_{tr} and stride s, resulting in a candidate set of windows $\mathcal{W}$. A subset $\mathcal{W}' \subset \mathcal{W}$ of size b is then selected to maximize the total pairwise dissimilarity:

$$\mathcal{W}' = \arg \max_{\mathcal{W}' \subset \mathcal{W}, |\mathcal{W}'|=b} \sum_{\{w^i, w^j\} \in \binom{\mathcal{W}'}{2}} \|w^i - w^j\|_2^2, \quad (3)$$

where each w^i is a sub-trajectory of length w_{tr}. Within each sampled window, we update the parameters θ_t using a gradient step on ℓ_{self}, then predict task progress from the updated representation, as described in 2.2.

3. Experiments

3.1. Datasets

We train and evaluate our model on expert visual trajectories paired with natural language task descriptions from the BridgeData V2 dataset (Walke et al., 2023), which spans a wide range of manipulation tasks, environments, and robot embodiments. For training, we use a curated subset consisting of 2,986 expert demonstrations covering pick-and-place manipulation tasks, but it does not include folding, sweeping, or stacking tasks. All demonstrations are collected using a single robot embodiment (WidowX 250) across four configurations of the ToyKitchen environment. An additional 287 expert demonstrations are held out as the in-distribution test set, referred to as `tk_pnp`.

We evaluate the ability of progress estimators to generalize across novel environments, tasks, and robot embodiments using subsets of BridgeData V2 curated to introduce variation along these axes. Environment shifts involve changes in scene layout or background. For example, `lm_pnp` is a pick-and-place task in front of a laundry machine, while `td_fold`, `ft_fold`, and `rd_fold` feature cloth folding on different surfaces. `ms_sweep` introduces a long-horizon sweeping task in a confined tray. Embodiment shifts are evaluated using the DeepThought robot, which differs from the training embodiment (WidowX 250) in both morphology and camera perspective. `dt_tk_pnp` (pick-and-place) and `dt_tk_stack` (stacking) retain the in-distribution environment with a new embodiment, while `dt_ft_stack` (stacking) and `dt_rd_pnp` (drawer pick-and-place) involve both embodiment and environment shifts. Each dataset includes 200 expert trajectories with task descriptions, except for `ms_sweep`, which contains 100 due to its size limitation. Appendix A provides a full description of our datasets.

3.2. Evaluation Metric

We evaluate the performance of a progress estimator using the *Value Order Correlation* (VOC) metric (Ma et al., 2024), which measures the alignment between the predicted progress values and the chronological order of the frames in a visual trajectory. Formally, let $p_1, \ldots, p_T$ denote the predicted progress values for a trajectory of T frames, and let $r_t = t$ represent the temporal index. VOC is defined as the Spearman rank correlation ρ_s between the predicted values and the temporal indices.

3.3. Baselines

The CLIP baseline performs zero-shot progress estimation by computing cosine similarity between frozen CLIP frame embeddings and the task description (Mahmoudieh et al., 2022). VLM-RM is a regularized CLIP method that projects features along the direction from a generic reference prompt to the task prompt (Rocamonde et al., 2024). CLIP-FT is a trained supervised regression model using frozen CLIP fea-

Table 1: Validation VOC scores for task progress estimation under distribution shifts relative to the training distribution. **ID** = In-Distribution, **ES** = Environment Shift, **EM** = Embodiment Shift, **ES & EM** = Environment and Embodiment Shift.

SHIFT	DATASET	CLIP	VLM-RM	CLIP-FT	GVL-0S	GVL-1S	TTT-EX	TTT-IM
ID	TK_PNP	0.0380	0.0289	0.2513	0.2691	0.2520	0.1957	**0.7822**
ES	LM_PNP	0.0170	0.0331	0.1489	0.3053	0.2719	0.1826	**0.7246**
	TD_FOLD	0.0310	0.0718	0.1521	0.3257	0.3181	0.1322	**0.7085**
	FT_FOLD	0.1081	0.0994	0.1619	0.3311	0.3874	0.1683	**0.6583**
	RD_FOLD	0.0952	0.0548	0.1257	0.3720	0.4057	0.1154	**0.6056**
	MS_SWEEP	-0.1285	-0.2260	0.1476	0.1580	0.1495	0.0102	**0.4898**
EM	DT_TK_PNP	0.0423	-0.0406	0.1490	0.2577	0.2151	0.2069	**0.8203**
	DT_TK_STACK	0.0347	0.0461	0.0991	0.2542	0.2770	0.0747	**0.7081**
ES & EM	DT_FT_STACK	0.0258	0.0282	0.0485	0.2122	0.2485	0.0439	**0.6979**
	DT_RD_PNP	0.0233	0.0408	0.2112	0.3292	0.3155	0.1559	**0.6951**

tures and a two-layer MLP head (Lin et al., 2022). GVL (Ma et al., 2024) is the state-of-the-art in-context learning approach that leverages autoregressive VLMs to predict task progress. In our experiments, GVL-0S refers to the zero-shot setting, where the model is prompted with only the test trajectory and task description. GVL-1S refers to a one-shot in-context setting, where a full shuffled trajectory from our training set, with its progress labels, is provided as an example. Both methods use the latest version of Gemini 1.5 Pro (Reid et al., 2024). Our method is evaluated in two adaptation configurations that differ in how they capture temporal context. The explicit-memory variant (TTT-EX) uses a fixed-length local context of the most recent frames ($k = 7$) and resets the adaptation module at each timestep, limiting adaptation to short-term dependencies. The implicit-memory variant (TTT-IM) updates the adaptation module sequentially ($k = 0$) without reset, enabling retention of past frames of the trajectory. Appendix B provides details about model architectures and training procedure.

3.4. Results

Our experiments show that the TTT-IM consistently outperforms both zero- and one-shot GVL, highlighting the importance of preserving temporal order in progress estimation. TTT-IM outperforms TTT-EX, suggesting that retaining memory of the test trajectory is more effective than relying on local context. We provide an additional analysis in Appendix C, demonstrating the impact of step-by-step compared to trajectory-level adaptation and the benefit of implicit memory. CLIP-FT and TTT-EX perform comparably to GVL on the in-distribution task but fail to generalize to out-of-distribution settings. Both CLIP and CLIP-Reg show limited performance, likely due to their lack of temporal modeling. For GVL-0S and GVL-1S, in some cases, in-context examples improve performance, while in others, they provide no benefit or even degrade performance.

Both GVL methods perform well on folding tasks (`td_fold`, `ft_fold`, `rd_fold`), but fail to achieve similar performance on stacking (`dt_tk_stack`, `dt_ft_stack`) and pick-and-place (`dt_tk_pnp`, `lm_pnp`) tasks, suggesting that the autoregressive VLM used in GVL may be biased toward folding-like robotic manipulations. In contrast, our test-time adaptation methods achieve consistent performance across all task types and distribution shifts, indicating better generalization than in-context learning. All methods, except CLIP-FT, perform worse on the long-horizon task `ms_sweep`, but TTT-IM achieves the highest VOC score. Under embodiment shifts, our method demonstrates strong generalization. In `dt_tk_pnp`, which uses a different robot embodiment but shares the same environment and tasks as the training set, both TTT-IM and TTT-EX even exceed their in-distribution performance, indicating that test-time adaptation can effectively transfer across robot embodiments. In contrast, GVL-1S performs poorly on both `dt_tk_pnp` and `dt_tk_stack`, suggesting that few-shot learning fails to generalize under embodiment shifts in our evaluation setup.

4. Discussion

Our results show that test-time adaptation enables progress estimators for robotic manipulation tasks to generalize across distribution shifts in task, environment, and embodiment. By updating the model incrementally at test time, our method captures both temporal and visual context, outperforming CLIP-based baselines, local-context adaptation, and autoregressive VLMs. This may be attributed to the fact that, despite their generalization capabilities through in-context learning, autoregressive VLMs are not explicitly trained for progress estimation, which requires temporal reasoning over visual trajectories. In contrast, our method requires access to expert visual trajectories during training, whereas in-context learning methods rely on large-scale pre-training.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Mensch, A., Millican, K., Reynolds, M., Borgeaud, S., Brock, A., et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Baumli, K., Baveja, S., Behbahani, F. M. P., Chan, H., Comanici, G., Flennerhag, S., Gazeau, M., Holsheimer, K., Horgan, D., Laskin, M., Lyle, C., Masoom, H., McKinney, K., Mnih, V., Neitz, A., Pardo, F., Parker-Holder, J., Quan, J., Rocktäschel, T., Sahni, H., Schaul, T., Schroecker, Y., Spencer, S., Steigerwald, R., Wang, L., and Zhang, L. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*, 2023. URL https://arxiv.org/abs/2312.09187.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, K. A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

Chen, A. S., Nair, S., and Finn, C. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

Dashora, N., Ghosh, D., and Levine, S. Viva: Video-trained value functions for guiding online rl from diverse data. *arXiv preprint arXiv:2503.18210*, 2025.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Ghosh, D., Walke, H. R., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., Luo, J., et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, 2024.

Ha, D. and Schmidhuber, J. World models. In *Advances in Neural Information Processing Systems*, pp. 2734–2742, 2018.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering atari with discrete world models. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Hughes, E., Dennis, M. D., Parker-Holder, J., Behbahani, F. M. P., Mavalankar, A., Shi, Y., Schaul, T., and Rocktäschel, T. Position: Open-endedness is essential for artificial superhuman intelligence. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=Bc4vZ2CX7E.

Hurst, O. A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mkadry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., drey Mishchenko, A., Baek, A., Jiang, A., toine Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C. L., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mély, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, P. D., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H.-B., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H. P., Ren, H., Chang, H., Chung, H. W., Kivlichan, I. D., O'Connell, I., Osband, I., Silber, I., Sohl, I., İbrahim Cihangir Okuyucu, Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J. W., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J. R., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik,

K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Long, O., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., laine Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., teusz Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M. A., Yang, M., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., Castro, M., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., talie Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N. A., Felix, N., Kudige, N., Keskar, N. S., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., ing Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R. H., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R. R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Xia, S., Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. Gpt-4o system card. *ArXiv*, abs/2410.21276, 2024. URL https://api.semanticscholar.org/CorpusID:273662196.

Lee, Y., Szot, A., Sun, S.-H., and Lim, J. J. Generalizable imitation learning from observation via inferring goal proximity. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16118–16130, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/868b7df964b1af24c8c0a9e43a330c6a-Abstract.html.

Lin, Z., Liu, H., Zisserman, A., and Vedaldi, A. Frozen clip models are efficient video learners. In *European Conference on Computer Vision (ECCV)*, 2022.

Ma, Y. J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., and Zhang, A. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Ma, Y. J., Kumar, V., Zhang, A., Bastani, O., and Jayaraman, D. LIV: Language-image representations and rewards for robotic control. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23301–23320. PMLR, 2023. URL https://proceedings.mlr.press/v202/ma23b.html.

Ma, Y. J., Hejna, J., Wahid, A., Fu, C., Shah, D., Liang, J., Xu, Z., Kirmani, S., Xu, P., Driess, D., Xiao, T., Tompson, J., Bastani, O., Jayaraman, D., Yu, W., Zhang, T., Sadigh, D., and Xia, F. Vision language models are in-context value learners. *arXiv preprint arXiv:2411.04549*, 2024. URL https://arxiv.org/abs/2411.04549.

Mahmoudieh, P., Goodman, N., and Finn, C. Zero-shot task transfer via goal-conditioned contrastive policy learning. In *International Conference on Machine Learning*, 2022.

Pătrăucean, V., Aytar, Y., Mottaghi, R., Rybkin, O., Lucic, M., Vondrick, C., Freeman, W. T., and Zisserman, A. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J.-B., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A. M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., jamin Lee, B., Viola, F., Reynolds, M., Xu, Y.,

Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K. W., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., Vilnis, L., car Chang, O., Morioka, N., Tucker, G., Zheng, C., Woodman, O., Attaluri, N., Kocisky, T., Eltyshev, E., Chen, X., Chung, T., Selo, V., Brahma, S., Georgiev, P., Slone, A., Zhu, Z., Lottes, J., Qiao, S., Caine, B., Riedel, S., Tomala, A., Chadwick, M., Love, J. C., Choy, P., Mittal, S., Houlsby, N., Tang, Y., Lamm, M., Bai, L., Zhang, Q., He, L., Cheng, Y., Humphreys, P., Li, Y., Brin, S., Cassirer, A., Miao, Y.-Q., Zilka, L., Tobin, T., Xu, K., Proleev, L., Sohn, D., Magni, A., Hendricks, L. A., Gao, I., Ontan'on, S., Bunyan, O., Byrd, N., Sharma, A., Zhang, B., Pinto, M., Sinha, R., Mehta, H., Jia, D., Caelles, S., Webson, A., Morris, A., Roelofs, B., Ding, Y., Strudel, R., Xiong, X., Ritter, M., Dehghani, M., Chaabouni, R., Karmarkar, A., Lai, G., Mentzer, F., Xu, B., Li, Y., Zhang, Y., Paine, T. L., Goldin, A., Neyshabur, B., Baumli, K., Levskaya, A., Laskin, M., Jia, W., Rae, J. W., Xiao, K., He, A., Giordano, S., Yagati, L., Lespiau, J.-B., Natsev, P., Ganapathy, S., Liu, F., Martins, D., Chen, N., Xu, Y., Barnes, M., May, R., Vezer, A., Oh, J., Franko, K., Bridgers, S., Zhao, R., Wu, B., Mustafa, B., Sechrist, S., Parisotto, E., Pillai, T. S., Larkin, C., Gu, C., Sorokin, C., Krikun, M., Guseynov, A., Landon, J., Datta, R., Pritzel, A., Thacker, P., Yang, F., Hui, K., Hauth, A., Yeh, C.-K., Barker, D., Mao-Jones, J., Austin, S., Sheahan, H., Schuh, P., Svensson, J., Jain, R., Ramasesh, V. V., Briukhov, A., Chung, D.-W., von Glehn, T., Butterfield, C., Jhakra, P., Wiethoff, M., Frye, J., Grimstad, J., Changpinyo, B., Lan, C. L., Bortsova, A., Wu, Y., Voigtlaender, P., Sainath, T. N., Smith, C., Hawkins, W., Cao, K., Besley, J., Srinivasan, S., Omernick, M., Gaffney, C., de Castro Surita, G., Burnell, R., Damoc, B., Ahn, J., Brock, A., Pajarskas, M., Petrushkina, A., Noury, S., Blanco, L., Swersky, K., Ahuja, A., Avrahami, T., Misra, V., de Liedekerke, R., Iinuma, M., Polozov, A., York, S., van den Driessche, G., Michel, P., Chiu, J., Blevins, R., Gleicher, Z., Recasens, A., Rrustemi, A., Gribovskaya, E., rko Roy, A., Gworek, W., Arnold, S. M. R., Lee, L., Lee-Thorp, J., Maggioni, M., Piqueras, E., Badola, K., Vikram, S., Gonzalez, L., Baddepudi, A., Senter, E., Devlin, J., Qin, J., Azzam, M., Trebacz, M., Polacek, M., Krishnakumar, K., yiin Chang, S., Tung, M., Penchev, I., Joshi, R., Olszewska, K., Muir, C., Wirth, M., Hartman, A. J., Newlan, J., Kashem, S., Bolina, V., Dabir, E., van Amersfoort, J. R., Ahmed, Z., Cobon-Kerr, J., Kamath, A. B., Hrafnkelsson, A. M., Hou, L., Mackinnon, I., Frechette, A., Noland, E., ance Si, X., Taropa, E., Li, D., Crone, P., Gulati, A., Cevey, S., Adler, J., Ma, A., Silver, D., Tokumine, S., Powell, R., Lee, S., Chang, M. B., Hassan, S., Mincu, D., Yang, A., Levine, N., Brennan, J., Wang, M., Hodkinson, S., Zhao, J., Lipschultz, J., Pope, A., Chang, M. B., Li, C., Shafey, L. E., Paganini, M., Douglas, S., Bohnet, B., Pardo, F., Odoom, S., Roșca, M., dos Santos, C. N., Soparkar, K., Guez, A., Hudson, T., Hansen, S., Asawaroengchai, C., Addanki, R., Yu, T., Stokowiec, W., Khan, M., Gilmer, J., Lee, J., Bostock, C. G., Rong, K., Caton, J., Pejman, P., Pavetic, F., Brown, G., Sharma, V., Luvci'c, M., mar Samuel, R., Djolonga, J., Mandhane, A., Sjosund, L. L., Buchatskaya, E., White, E., Clay, N., Jiang, J., Lim, H., Hemsley, R., Labanowski, J., Cao, N. D., Steiner, D., Hashemi, S. H., Austin, J., Gergely, A., Blyth, T., Stanton, J., Shivakumar, K., Siddhant, A., Andreassen, A., Araya, C. L., Sethi, N., Shivanna, R., Hand, S., Bapna, A., Khodaei, A., Miech, A., Tanzer, G., Swing, A., Thakoor, S., Pan, Z., Nado, Z., Winkler, S., Yu, D., Saleh, M., Maggiore, L., Barr, I., Giang, M., Kagohara, T., Danihelka, I., Marathe, A., Feinberg, V., Elhawaty, M., Ghelani, N., Horgan, D., Miller, H., Walker, L., Tanburn, R., Tariq, M., Shrivastava, D., Xia, F., Chiu, C.-C., Ashwood, Z., Baatarsukh, K., Samangooei, S., Alcober, F., Stjerngren, A., Komarek, P., Tsihlas, K., Boral, A., Comanescu, R., Chen, J., Liu, R., Bloxwich, D., Chen, C., Sun, Y., aoyu Feng, F., Mauger, M., Dotiwalla, X., Hellendoorn, V., Sharman, M., Zheng, I., Haridasan, K., Barth-Maron, G., Swanson, C., Rogozi'nska, D., Andreev, A., Rubenstein, P. K., Sang, R., Hurt, D., Elsayed, G., shen Wang, R., Lacey, D., Ili'c, A., Zhao, Y., Han, W., Aroyo, L., Iwuanyanwu, C., Nikolaev, V., Lakshminarayanan, B., Jazayeri, S., Kaufman, R. L., Varadarajan, M., Tekur, C., Fritz, D., Khalman, M., Reitter, D., Dasgupta, K., Sarcar, S., Ornduff, T., Snaider, J., Huot, F., Jia, J., Kemp, R., Trdin, N., Vijayakumar, A., Kim, L., Angermueller, C., Lao, L., Liu, T., Zhang, H., Engel, D., Greene, S., White, A., Austin, J., Taylor, L., Ashraf, S., Liu, D., Georgaki, M., Cai, I., Kulizhskaya, Y., Goenka, S., Saeta, B., Vodrahalli, K., Frank, C., de Cesare, D., Robenek, B., Richardson, H., moud Alnahlawi, M., pher Yew, C., Ponnapalli, P., Tagliasacchi, M., Korchemniy, A., Kim, Y., Li, D., Rosgen, B., Levin, K., Wiesner, J., Banzal, P., Srinivasan, P., Yu, H., cCauglar Unlu, Reid, D., Tung, Z., Finchelstein, D. F., Kumar, R., Elisseeff, A., Huang, J., Zhang, M., Zhu, R., Aguilar, R., Gim'enez, M., Xia, J., Dousse, O., Gierke, W., Yeganeh, S. H., Yates, D., Jalan, K., Li, L., Latorre-Chimoto, E., Nguyen, D. D., Durden, K., Kallakuri, P., Liu, Y., Johnson, M., Tsai, T., Talbert, A., Liu, J., Neitz, A., Elkind, C., Selvi, M., Jasarevic, M., Soares, L. B., Soares, L. B., Wang, P., Wang, A. W., Ye, X., Kallarackal, K., Loher, L., Lam, H., Broder, J., Holtmann-Rice, D. N., Martin, N., Ramadhana, B., Toyama, D., Shukla, M., Basu, S., Mohan, A., Fernando, N., Fiedel, N., Paterson, K., Li, H., Garg, A., Park, J., Choi, D., Wu, D., Singh, S., Zhang, Z., Globerson, A., Yu, L., Carpenter, J., de Chau-

mont Quitry, F., Radebaugh, C., Lin, C.-C., Tudor, A., Shroff, P., Garmon, D., Du, D., Vats, N., Lu, H., Iqbal, S., Yakubovich, A., Tripuraneni, N., Manyika, J., roon Qureshi, H., Hua, N., Ngani, C., Raad, M. A., Forbes, H., Bulanova, A., Stanway, J., Sundararajan, M., Ungureanu, V., Bishop, C., Li, Y., Venkatraman, B., Li, B., Thornton, C., Scellato, S., Gupta, N., Wang, Y., Tenney, I., Wu, X., Shenoy, A., Carvajal, G., Wright, D. G., Bariach, B., Xiao, Z., Hawkins, P., Dalmia, S., Farabet, C., Valenzuela, P., Yuan, Q., Welty, C. A., Agarwal, A., Chen, M., Kim, W., Hulse, B., Dukkipati, N., Paszke, A., Bolt, A., Davoodi, E., Choo, K., Beattie, J., Prendki, J., Vashisht, H., beca Santamaria-Fernandez, R., Cobo, L. C., Wilkiewicz, J., Madras, D., Elqursh, A., Uy, G., Ramirez, K., Harvey, M., Liechty, T., Zen, H., Seibert, J., Hu, C. H., Khorlin, A. Y., Le, M., Aharoni, A., Li, M., Wang, L., Kumar, S., Lince, A., Casagrande, N., Hoover, J., Badawy, D. E., Soergel, D., Vnukov, D., Miecnikowski, M., Simsa, J., Koop, A., Kumar, P., Sellam, T., Vlasic, D., Daruki, S., Shabat, N., Zhang, J., Su, G., Krishna, K., Zhang, J., Liu, J., Sun, Y., Palmer, E., Ghaffarkhah, A., Xiong, X., Cotruta, V., Fink, M., Dixon, L., Sreevatsa, A., Goedeckemeyer, A., Dimitriev, A., Jafari, M., Crocker, R., Fitzgerald, N., Kumar, A., Ghemawat, S., Philips, I., Liu, F., Liang, Y., Sterneck, R., Repina, A., Wu, M., Knight, L., Georgiev, M., Lee, H., Askham, H., Chakladar, A., Louis, A., Crous, C., Cate, H., Petrova, D., Quinn, M., Owusu-Afriyie, D., Singhal, A., Wei, N., Kim, S., Vincent, D., Nasr, M., Shumailov, I., Choquette-Choo, C. A., Tojo, R., Lu, S., de Las Casas, D., Cheng, Y., Bolukbasi, T., ine Lee, K., Fatehi, S., Ananthanarayanan, R., Patel, M., Kaed, C. E., Li, J., Sygnowski, J., Belle, S. R., Chen, Z., Konzelmann, J., Põder, S., Garg, R., Koverkathu, V., Brown, A., Dyer, C., Liu, R., Nova, A., Xu, J., Bai, J., Petrov, S., Hassabis, D., Kavukcuoglu, K., Dean, J., Vinyals, O., and Chronopoulou, A. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024. URL https://api.semanticscholar.org/CorpusID:268297180.

Rocamonde, J., Montesinos, V., Nava, E., Perez, E., and Lindner, D. Vision-language models are zero-shot reward models for reinforcement learning. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=N0I2RtD8je.

Silver, D. and Sutton, R. S. Welcome to the era of experience. Technical report, Google AI, 2025.

Singh, A. R., Kumar, V., and Gupta, A. Generalizable imitation learning from observation via inferring goal proximity. In *NeurIPS*, 2021.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020. URL http://proceedings.mlr.press/v119/sun20b.html.

Sun, Y., Li, X., Dalal, K., Xu, J., Vikram, A., Zhang, G., Dubois, Y., Chen, X., Wang, X., Koyejo, S., Hashimoto, T., and Guestrin, C. Learning to (learn at test time): Rnns with expressive hidden states. *CoRR*, abs/2407.04620, 2024. doi: 10.48550/arXiv.2407.04620. URL https://arxiv.org/abs/2407.04620.

Team, S., Raad, M. A., Ahuja, A., Barros, C., Besse, F., Bolt, A., Bolton, A., Brownfield, B., Buttimore, G., Cant, M., Chakera, S., Chan, S. C. Y., Clune, J., Collister, A., Copeman, V., Cullum, A., Dasgupta, I., de Cesare, D., Trapani, J. D., Donchev, Y., Dunleavy, E., Engelcke, M., Faulkner, R., Garcia, F., Gbadamosi, C., Gong, Z., Gonzalez, L., Gupta, K., Gregor, K., Hallingstad, A. O., Harley, T., Haves, S., Hill, F., Hirst, E., Hudson, D. A., Hudson, J., Hughes-Fitt, S., Rezende, D. J., Jasarevic, M., Kampis, L., Ke, N. R., Keck, T., Kim, J., Knagg, O., Kopparapu, K., Lampinen, A. K., Legg, S., Lerchner, A., Limont, M., Liu, Y., Loks-Thompson, M., Marino, J., Cussons, K. M., Matthey, L., Mcloughlin, S., Mendolicchio, P., Merzic, H., Mitenkova, A., Moufarek, A., Oliveira, V., Oliveira, Y. G., Openshaw, H., Pan, R., Pappu, A., Platonov, A., Purkiss, O., Reichert, D. P., Reid, J., Richemond, P. H., Roberts, T., Ruscoe, G., Elias, J. S., Sandars, T., Sawyer, D. P., Scholtes, T., Simmons, G., Slater, D., Soyer, H., Strathmann, H., Stys, P., Tam, A. C., Teplyashin, D., Terzi, T., Vercelli, D., Vujatovic, B., Wainwright, M., Wang, J. X., Wang, Z., Wierstra, D., Williams, D., Wong, N., York, S., and Young, N. Scaling instructable agents across many simulated worlds. *CoRR*, abs/2404.10179, 2024. doi: 10.48550/arXiv.2404.10179. URL https://arxiv.org/abs/2404.10179.

Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A. W., Myers, V., Kim, M. J., Du, M., Lee, A., Fang, K., Finn, C., and Levine, S. Bridgedata v2: A dataset for robot learning at scale. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

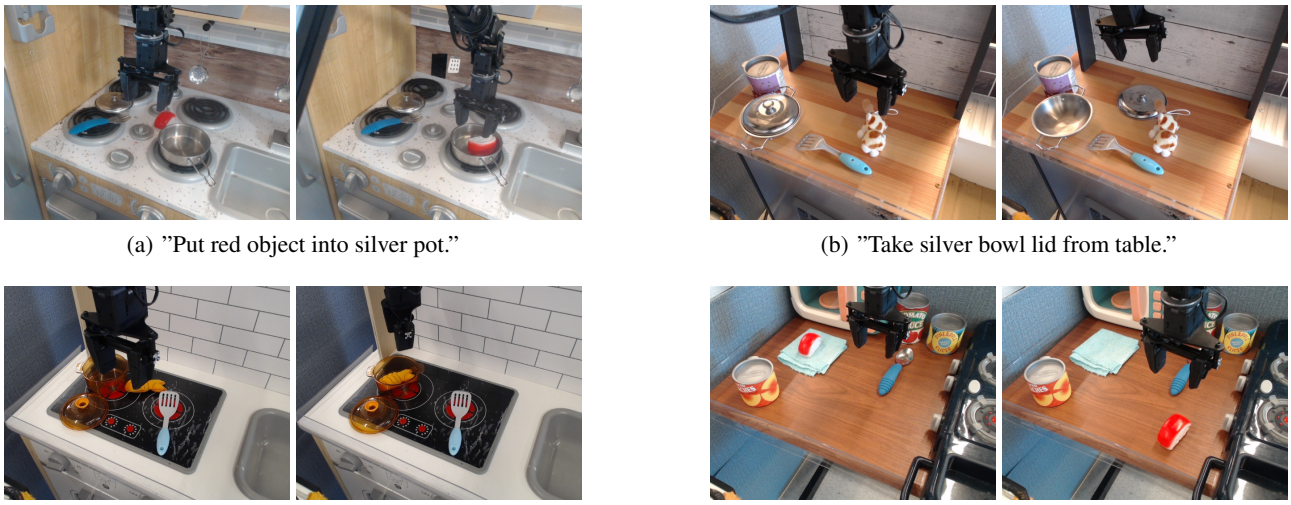
Wang, R., Sun, Y., Gandelsman, Y., Chen, X., Efros, A. A., and Wang, X. Test-time training on video streams. *CoRR*, abs/2307.05014, 2023. doi: 10.48550/arXiv.2307.05014. URL https://arxiv.org/abs/2307.05014.

Xu, Y., Wu, J., Zhang, Y., Qiu, M., and Wang, Y. Adapting to

length shift: Flexilength network for trajectory prediction. In *CVPR*, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A. Training Dataset



(a) "Put red object into silver pot."

(b) "Take silver bowl lid from table."

(c) "Place yellow object inside pot."

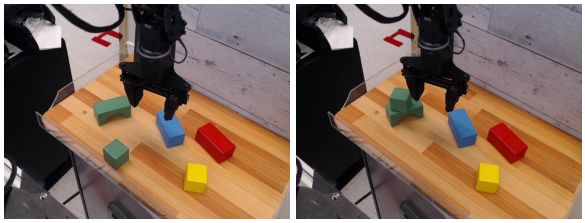
(d) "Move red object from pot to left burner."

Figure 2: Each subfigure shows the start and end frames from an expert demonstration used for training, along with its natural language task description. Demonstrations are collected across four distinct ToyKitchen environments.

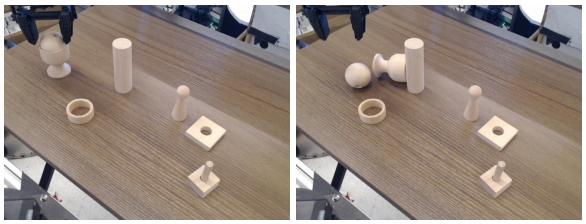
Table 2: Dataset descriptions with task type, environment, and shift breakdown relative to our training distribution. Checkmarks indicate a distribution shift along the task, environment, or embodiment dimension.

Dataset	Task Type	Environment	Task	Environment	Embodiment
`tk_pnp`	pick-and-place	toy kitchen			
`lm_pnp`	pick-and-place	laundry machine		✓	
`td_fold`	fold cloth	tabletop (dark wood)	✓	✓	
`ft_fold`	fold cloth	folding table	✓	✓	
`rd_fold`	fold cloth	robot desk	✓	✓	
`ms_sweep`	sweep	folding table (tray)	✓	✓	
`dt_tk_pnp`	pick-and-place	toy kitchen			✓
`dt_tk_stack`	stack blocks	toy kitchen	✓		✓
`dt_ft_stack`	stack blocks	folding table	✓	✓	✓
`dt_rd_pnp`	pick-and-place	robot desk (drawer)		✓	✓

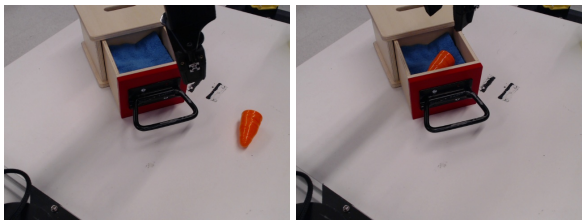
(a) `dt_tk_ms`: "Move yellow object from burner to center."



(b) `dt_tk_stack`: "Place green cube on top of arch."

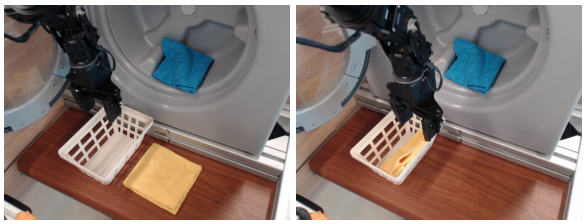


(c) `dt_ft_stack`: "Move egg to table."

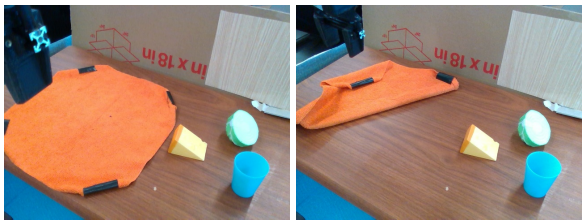


(d) `dt_rd_pnp`: "Put orange object in drawer."

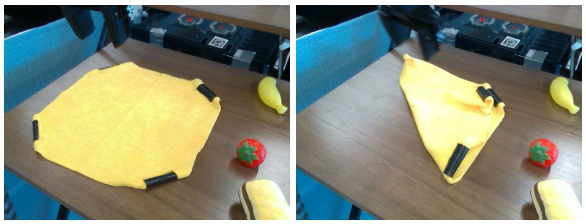
Figure 3: Each subfigure shows the start and end frames from an evaluation trajectory under embodiment shift, along with its natural language task description. The top row depicts tasks in the same environment (ToyKitchen) using a different robot (DeepThought), while the bottom row includes tasks that also involve new environments.



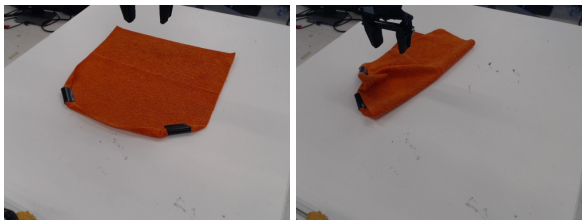
(a) `lm_pnp`: "Place blue cloth inside washer."



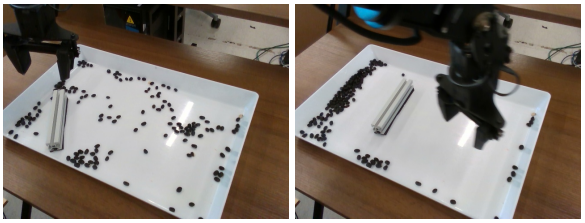
(b) `td_fold`: "Fold cloth bottom right to top left."



(c) `ft_fold`: "Fold clothes from left to right."



(d) `rd_fold`: "Fold cloth bottom right to top left."



(e) `ms_sweep`: "Sweep into pile."

Figure 4: Each subfigure shows the start and end frames from a trajectory under environment shift, along with its natural language task description. All tasks are performed using the same robot embodiment across different environments.

B. Model Architecture

B.1. Model and Training Details

We use the OpenCLIP ViT-B/32 encoder pre-trained on the OpenAI dataset as a frozen backbone. Each visual observation and task description is encoded using the CLIP vision and text encoders to produce their representation, respectively. We opted for a joint (concatenation-based) representation over an element-wise product based on improved validation performance. The adaptation module f_{adapt} is a two-layer residual MLP with GELU activation and a projection dimension $d' = 64$. The model is trained using the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and a cosine learning rate schedule with 10% warmup. We use a batch size of 32 and pad all trajectories to a maximum length of 120 frames (matching the longest sequence in the training set). The weighting coefficient λ_{self} of the self-supervised loss in the total training objective is set to 0.5, selected based on validation performance. We train for 5 epochs, as validation VOC typically plateaus early, with extended training providing no further improvement. For dissimilarity-based sampling during training, $w_{\text{train}} = 8$ and `b` = 8 are chosen after validation. All experiments were run on NVIDIA RTX 6000 Ada Generation GPUs.

B.2. Test-Time Training Hyperparameters

At inference time, we adapt only the temporal adaptation module f_{adapt}, using the same projection dimension ($d' = 64$) as in training. We perform adaptation over a single gradient step (`t_ep` = 1), using a learning rate of 0.1. This configuration was selected from a sweep over learning rates $\{5.0, 1.0, 0.1, 0.01\}$ and adaptation steps $\{1, 5, 10\}$, based on performance on the validation set. Unless otherwise noted, the adaptation module is not reset between evaluation episodes.

B.3. Baselines

Both TTT-IM and TTT-EX use `t_ep` = 1, projection dimension $d' = 64$, and learning rates η of 0.1 and 1.0, respectively, selected via hyperparameter tuning. TTT-IM uses no sliding window during inference ($k = 0$), whereas TTT-IM uses a sliding window of size 8 ($k = 7$). CLIP-FT shares the same architecture as our method but excludes the adaptation module and self-supervised loss. It uses a frozen CLIP encoder, followed by a linear projection and a two-layer MLP to predict task progress. The model is trained with standard supervised regression, without meta-learning or test-time adaptation. To increase expressivity, it uses an $8\times$ larger projection matrix and $10\times$ more training steps than our method (Lin et al., 2022). For VLM-RM, we use a baseline prompt that describes the environment. For both GVL-0S and GVL-1S, we use the latest version of Gemini 1.5 Pro, `gemini-1.5-pro-latest` (Reid et al., 2024), whereas the original GVL implementation used an earlier release, `gemini-1.5-pro`. We also evaluated GVL using the open-source autoregressive VLM `Qwen-VL 2.5` (Yang et al., 2024), which failed to overcome temporal bias despite frame shuffling, while GPT-4o (Hurst et al., 2024) consistently declined to produce scalar progress estimates in our setup. All evaluations followed the prompt template introduced in the original GVL work (Ma et al., 2024).

C. Trajectory-Level vs. Step-by-Step Adaptation

Table 3: Validation VOC scores for progress estimation under distribution shifts. ID = In-Distribution, ES = Environment Shift, EM = Embodiment Shift, ES+EM = Environment and Embodiment Shift.

Shift	Dataset	TTT-TR	TTT-RS	TTT-IM
ID	tk_pnp	0.1917	0.1918	**0.7822**
ES	lm_sweep	0.1843	0.1854	**0.7246**
	td_fold	0.1402	0.1393	**0.7085**
	ft_fold	0.1302	0.1311	**0.6583**
	rd_fold	0.1161	0.1172	**0.6056**
	ms_ft_sweep	-0.0225	-0.0191	**0.4898**
EM	dt_tk_pnp	0.2123	0.2118	**0.8203**
	dt_tk_stack	0.0833	0.0830	**0.7081**
ES+EM	dt_ft_stack	0.0558	0.0537	**0.6979**
	dt_rd_pnp	0.1665	0.1660	**0.6951**

We evaluate three adaptation strategies: (1) TTT-TR, which updates the adaptation module once per trajectory; (2) TTT-RS, which resets the adaptation module at every step and updates using only the current visual observation; and (3) TTT-IM, our implicit memory variant that incrementally updates the adaptation parameters across timesteps. All approaches use the same architecture and initialization. As shown in Table 3, TTT-TR and TTT-RS achieve nearly identical results, indicating that applying a single update per trajectory offers no advantage over resetting at each step. In contrast, TTT-IM consistently outperforms TTT-TR and TTT-RS, highlighting the benefit of retaining temporal context throughout the trajectory.