Supplementary Material

Guide Your Agent with Adaptive Multimodal Rewards

A Experiment Details

In this section, we describe the details for implementing Multimodal Reward Decision Transformer, and we provide our source code in the supplementary material.

Procgen details We utilize a publicly available implementation[3] to replicate the environments introduced by Di Langosco et al. [9]. We modify the simulator of the environments to render higher-resolution images to leverage pre-trained multimodal representations for both our method and baselines. In this particular setup, the observations obtained from the environment at each timestep t comprise an RGB image with dimensions of $256 \times 256 \times 3$ and a natural language instruction that delineates the desired goal. Throughout our experiments, we adhere to the *hard environment difficulty* as described in [7]. Maximum episode length for all tasks is 500. To gather expert demonstrations used for training data, we train PPG [8] agents on 500 training levels for 200M timesteps per task using hyperparameters provided in Cobbe et al. [8]. For evaluation purposes, we assess the test performance on 1000 different levels, encompassing previously unseen themes and goals that differ from those employed in training.

Architecture details Both *Instruct*RL (Liu et al., 2022) and MRDT employ ViT-B/16 as the transformer-policy and pre-trained multimodal transformer encoder (M3AE; [12]) in all experiments, unless stated otherwise. Inspired by Gao et al. [11], we attach an additional 2-layer MLP to the end of a pre-trained multimodal transformer encoder and perform residual-style feature blending with the pre-trained features. In the training phase, we apply gradients only to the weight of these linear layers. Through empirical evaluation, we observe that this architecture yields superior performance in both our method and the baseline.

Training details We use $256 \times 256 \times 3$ RGB observations for training the return-conditioned policy. To stabilize training, we normalize multimodal returns following the method proposed by Chen et al. (2021), dividing them by 1000 in all experiments. We use the AdamW optimizer (Loshchilov et al., 2018) with a learning rate of 5×10^{-4} and weight decay 5×10^{-5}. A cosine decay schedule is utilized to adjust the training learning rate. In CoinRun experiments, data augmentation techniques such as color jitter and random rotation are applied to the RGB images o_t while maintaining alignment in the context. However, no augmentation is applied to RGB images in Maze I/II experiments. For scaling the return prediction loss in training the return-conditioned policy, we set $\lambda = 0.01$ in CoinRun experiments and $\lambda = 0.001$ in Maze I/II experiments. During the fine-tuning of the pre-trained multimodal encoder, a 2-layer MLP is attached to the end of both CLIP image and text encoders. Additionally, an extra 2-layer MLP is added as an action prediction layer for the IDM objective. The model is trained for 20 epochs, and the one with the lowest validation loss is used for generating multimodal rewards. To scale the IDM loss in fine-tuning CLIP, we employ $\beta = 1.5$ in CoinRun experiments and $\beta = 2.0$ in Maze I/II experiments.

Computation We use 24 CPU cores (Intel Xeon CPU @ 2.2GHz) and 2 GPUs (NVIDIA A100 40GB GPU) for training return-conditioned policy. The training of MRDT for 50 epochs takes approximately 4 hours for CoinRun experiments with the largest dataset size. For fine-tuning CLIP, we use 24 CPU cores (Intel Xeon CPU @ 2.2GHz) and 1 GPU (NVIDIA A100 40GB GPU), and it takes approximately 1.5 hours for Coinrun experiments.

[3] https://github.com/JacobPfau/procgenAISC

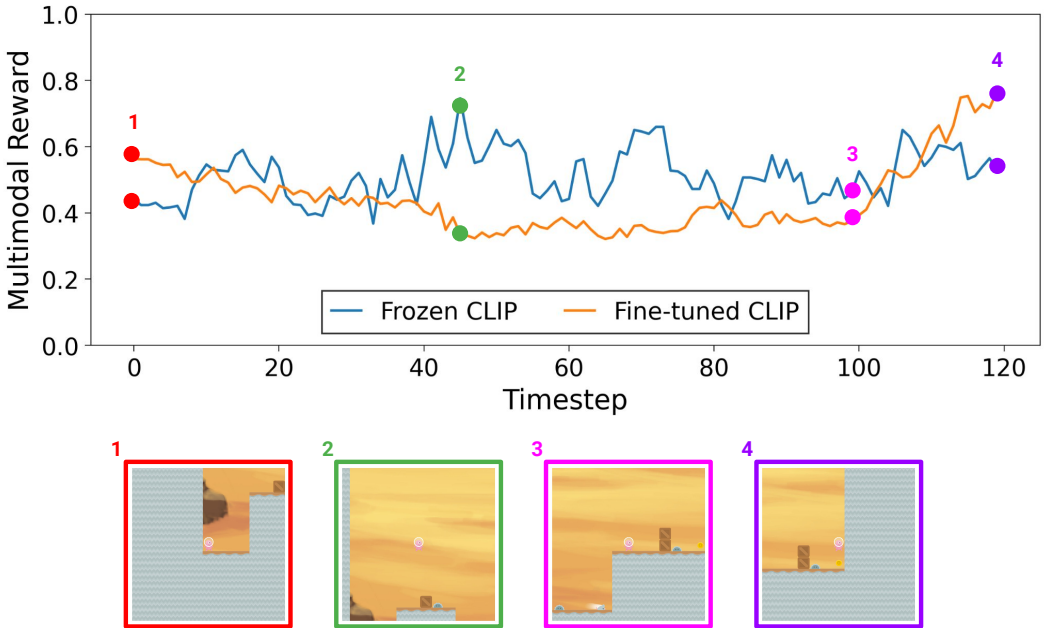
Hypeparameters We report the hyperparameters used in our experiments in Table 3.

Table 3: Hyperparameters of Multimodal Reward Decision Transformer (MRDT). Unless specified, we use the same hyperparameters used in *Instruct*RL [16].

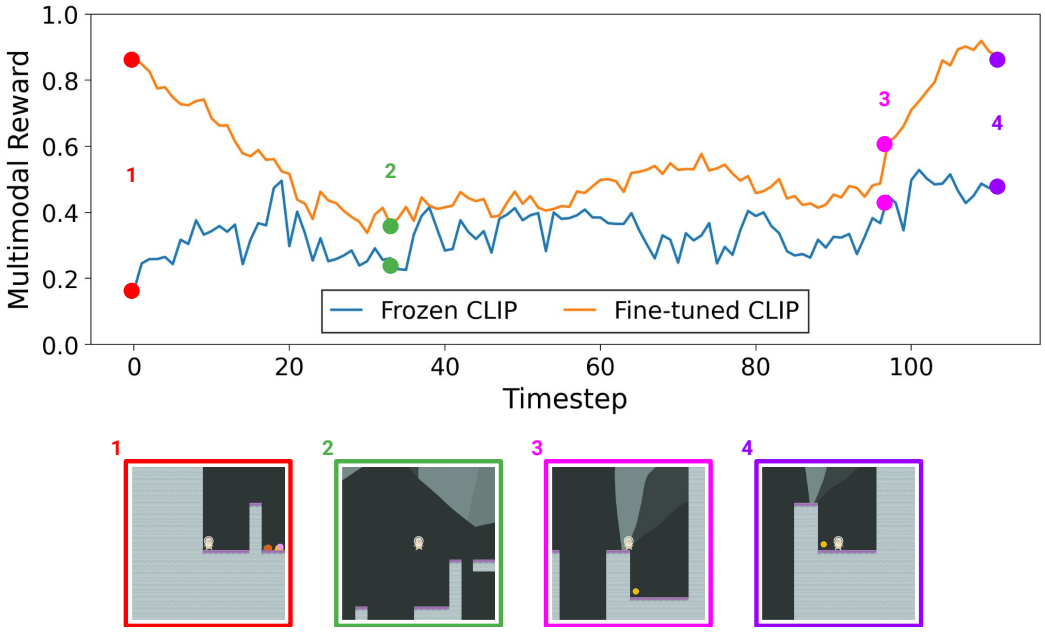
Hyperparameter	**Value**
Policy batch size	64
Policy epochs	50
Policy context length	4
Policy learning rate	0.0005
Policy optimizer	AdamW [18]
Policy optimizer momentum	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Policy weight decay	0.00005
Policy learning rate decay	Linear warmup and cosine decay (see code for details)
Policy context length	4
Policy transformer size	2 layers, 4 heads, 768 units
Fine-tuned CLIP batch size	64
Fine-tuned CLIP epochs	20
Fine-tuned CLIP learning rate	0.0001
Fine-tuned CLIP weight decay	0.001
Fine-tuned CLIP adapter layer size	2 layers, 1024 units
Fine-tuned CLIP optimizer	AdamW [18]
Fine-tuned CLIP optimizer momentum	$\beta_1 = 0.9$, $\beta_2 = 0.999$

B Qualitative Results of Multimodal Rewards

In Figure 9, 10, 11, we present the curves of multimodal rewards for frozen/fine-tuned CLIP on the trajectories from training/held-out evaluation environments. We find that the multimodal reward exhibits an overall increasing trend as the agent approaches the goal in both frozen and fine-tuned CLIP, irrespective of the training and held-out evaluation environments. Furthermore, we observe that fine-tuned CLIP not only induces a reward that is temporally smoother in the intermediate stages compared to frozen CLIP (see Figure 9) but also demonstrates a steeper upward reward curve (see Figure 10, 11). These results support the claim that the quality of multimodal rewards from the fine-tuned CLIP outperforms those from the frozen CLIP (Section 4.2). Video examples of the trajectories are provided in the supplementary material.

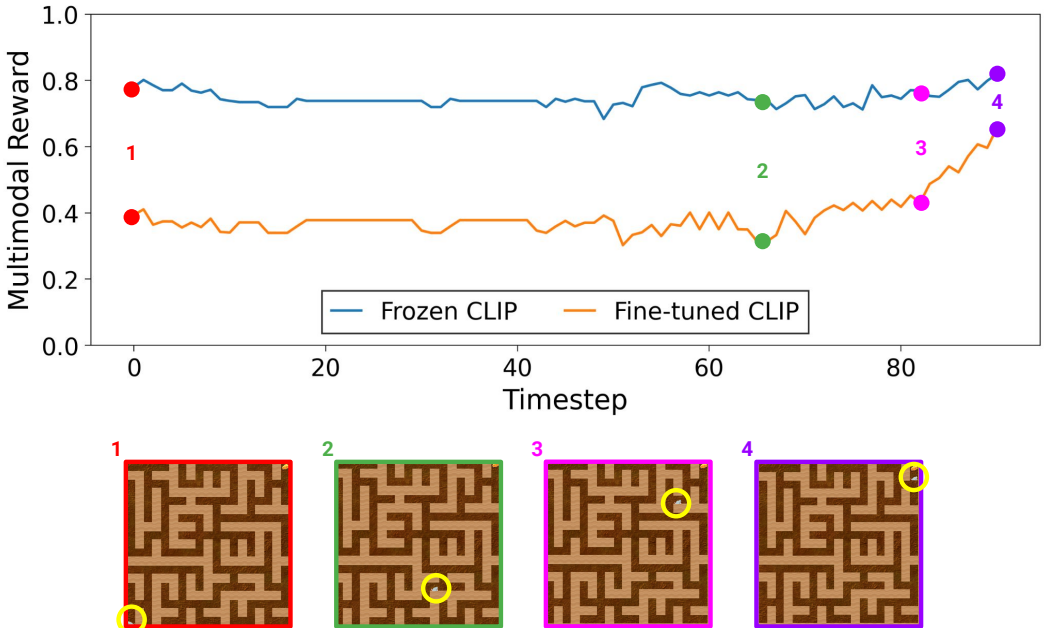


(a) Multimodal reward curve on the training environment.

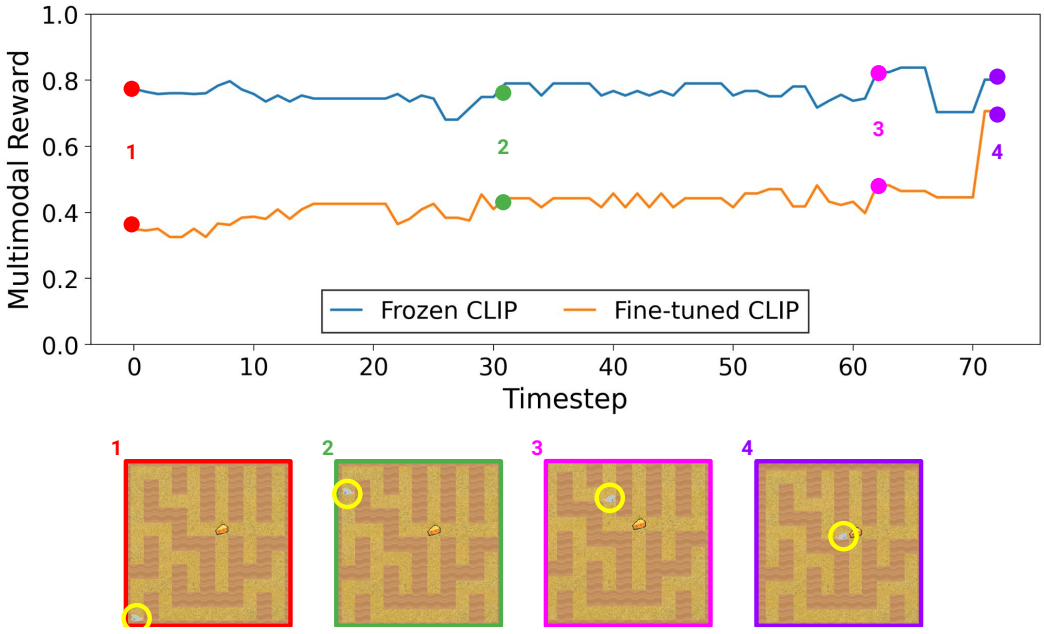


(b) Multimodal reward curve on the held-out evaluation environment.

Figure 9: Qualitative results of multimodal rewards in CoinRun environments.

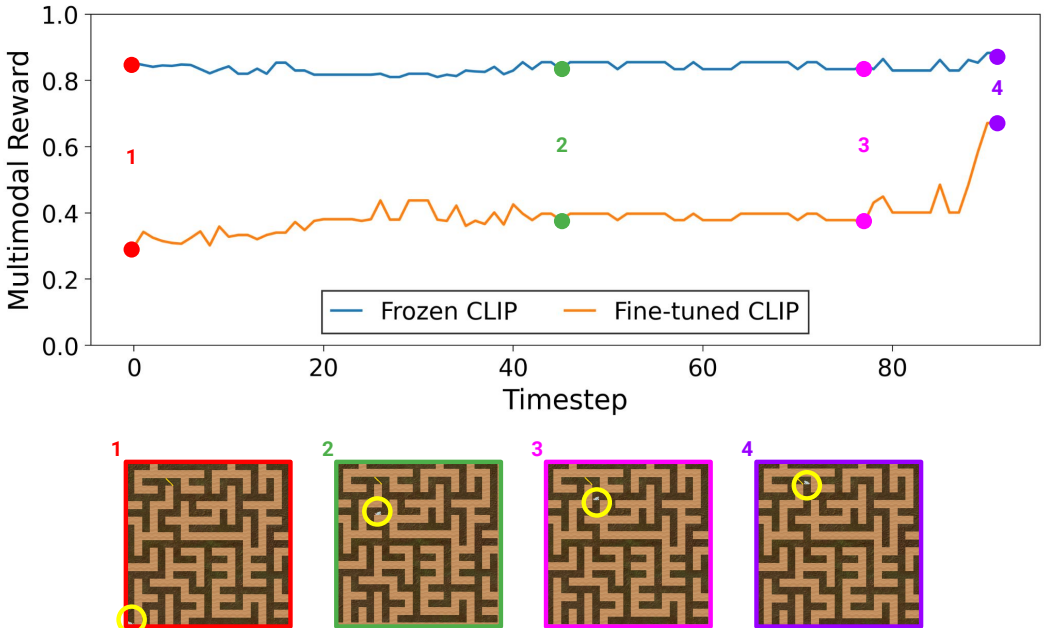


(a) Multimodal reward curve on the training environment.

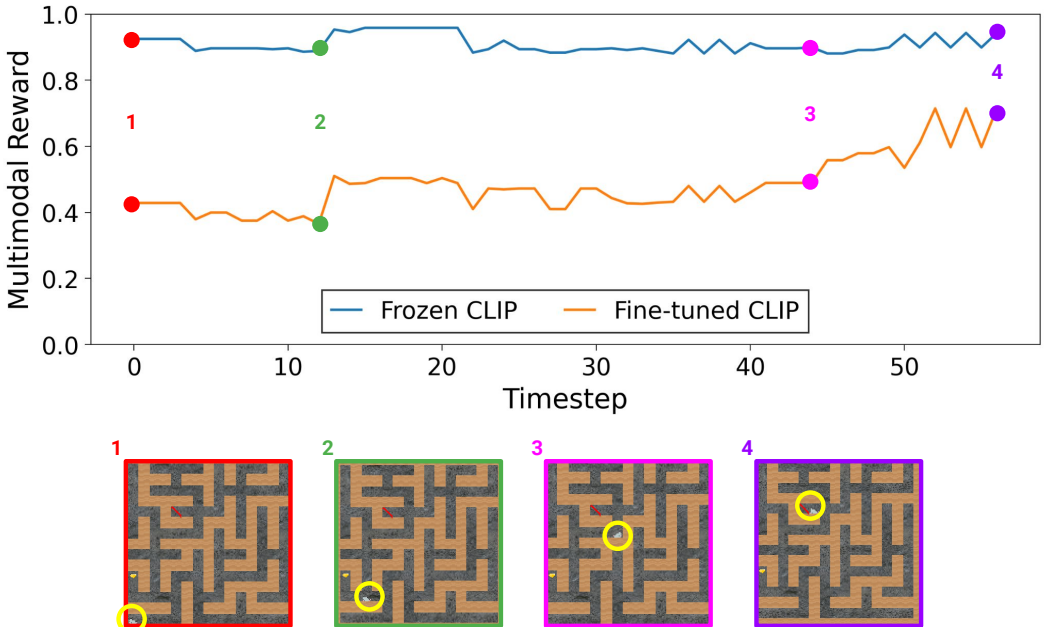


(b) Multimodal reward curve on the held-out evaluation environment.

Figure 10: Qualitative results of multimodal rewards in Maze I environments.



(a) Multimodal reward curve on the training environment.



(b) Multimodal reward curve on the held-out evaluation environment.

Figure 11: Qualitative results of multimodal rewards in Maze II environments.

C Additional Experiments

Table 4: Expert-normalized scores on training/evaluation CoinRun environments investigating the effect of hyperparameter λ adjusting the scale of return prediction loss in training return-conditioned policy. The result shows the mean and standard variation averaged over three runs.

λ	$\mathcal{L}_{\text{FT}}$	Train (%)	Eval (%)
0.001	✗	89.93% ± 3.94%	62.65% ± 10.12%
	✓	85.42% ± 1.80%	71.69% ± 5.71%
0.01	✗	89.58% ± 2.08%	63.32% ± 2.01%
	✓	90.28% ± 1.59%	72.36% ± 3.48%
0.1	✗	87.15% ± 2.62%	62.65% ± 10.31%
	✓	85.76% ± 3.18%	73.37% ± 3.48%
1.0	✗	87.15% ± 4.70%	61.64% ± 6.38%
	✓	81.25% ± 1.04%	73.37% ± 2.66%

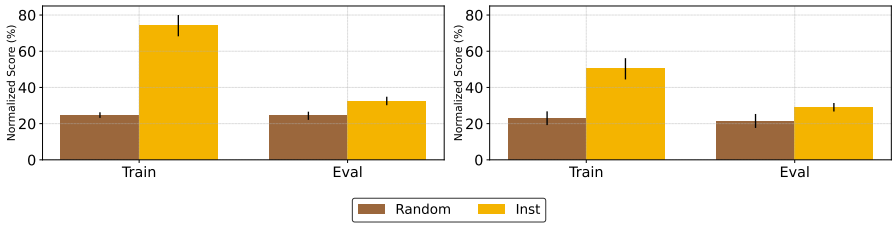


Figure 12: Expert-normalized scores on training/evaluation environments of MRDT trained using multimodal rewards generated with (i) instructive text (*i.e.,* Inst) and (ii) random text (*i.e.,* Random) in Maze I environments (left) and Maze II environments (right). The result shows the mean and standard deviation averaged over three runs.

Effect of scaling return prediction loss We investigate how the coefficient λ, which determines the weight of the return prediction loss in training return-conditioned policy, affects the performance of MRDT. To this end, we test various values of λ in CoinRun environments. Table 4 shows the performance of MRDT in training/held-out evaluation environments with different λ. We find that performance is not significantly different according to the value of λ in the held-out evaluation environments. These results indicate that MRDT is robust to the choice of hyperparameter λ.

Extra ablation study on text instructions In Figure 12, we further investigate whether MRDT leverages adaptive signals from multimodal rewards in decision-making. We evaluate the quality of rewards generated with instructive text (*i.e.,* Inst) and random text (*i.e.,* Random) in Maze I/II environments. Specifically, we use a natural language instruction for each environment, as described in Section 4 for Inst, and "NeurIPS 2023 will be held again at the New Orleans Ernest N. Morial Convention Center" for Random. We find that using random text instructions results in a decline in performance in both training and evaluation environments. These findings align with the trend observed in Figure 8.

D Limitation and Future Work

One limitation of our work is that we currently rely on a single image-text pair to compute the multimodal reward at every timestep t. Although our approach has shown effectiveness both quantitatively and qualitatively, there are tasks where rewards depend on the history of past observations (*i.e.,* non-Markovian) [2, 3, 15]. To address this limitation, it would be valuable to explore the extension of our method to incorporate video-text pairs for calculating multimodal rewards. This extension could involve generating multimodal rewards using pre-trained video-text multimodal representations [17, 19, 24, 21], which presents an intriguing avenue for better generalization across various goals in behavior learning. Another aspect to consider is that the tasks we have examined so far are relatively simple, as they involve only a single condition for success. To tackle more complex problems, we are interested in investigating approaches that leverage large language models [13, 14, 1, 10] in conjunction with our method. Finally, an interesting direction to explore would be the utilization of multimodal rewards in combination with extrinsic rewards [22, 20, 5].

E Potential Negative Societal Impacts

We do not anticipate significant negative societal impacts in that our method is now limited to playing simple simulation games. However, if our method is applied in real-world scenarios, privacy concerns may arise considering that behavior cloning agents used in such applications, like autonomous driving [23] or real-time control [4, 10], require large amounts of data, which often contain controversial information. Additionally, a behavior cloning policy presents a challenge as it imitates specified demonstrations, potentially including undesirable actions. If some bad actions are included in expert demonstrations (*e.g.,* behaviors that may be violent or harmful to the pedestrians are contained in the training data for mobile manipulation tasks), the policy could have significant negative impacts on users. To address this concern, future directions should focus on developing agents with safe adaptation in addition to performance enhancement efforts.

References

[1] Ahn, Michael, Brohan, Anthony, Brown, Noah, Chebotar, Yevgen, Cortes, Omar, David, Byron, Finn, Chelsea, Gopalakrishnan, Keerthana, Hausman, Karol, Herzog, Alex, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[2] Bacchus, Fahiem, Boutilier, Craig, and Grove, Adam. Rewarding behaviors. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1160–1167, 1996.

[3] Bacchus, Fahiem, Boutilier, Craig, and Grove, Adam. Structured solution methods for non-markovian decision processes. In *AAAI/IAAI*, pp. 112–117. Citeseer, 1997.

[4] Brohan, Anthony, Brown, Noah, Carbajal, Justice, Chebotar, Yevgen, Dabis, Joseph, Finn, Chelsea, Gopalakrishnan, Keerthana, Hausman, Karol, Herzog, Alex, Hsu, Jasmine, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[5] Burda, Yuri, Edwards, Harrison, Storkey, Amos, and Klimov, Oleg. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

[6] Chen, Lili, Lu, Kevin, Rajeswaran, Aravind, Lee, Kimin, Grover, Aditya, Laskin, Misha, Abbeel, Pieter, Srinivas, Aravind, and Mordatch, Igor. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[7] Cobbe, Karl, Hesse, Chris, Hilton, Jacob, and Schulman, John. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

[8] Cobbe, Karl W, Hilton, Jacob, Klimov, Oleg, and Schulman, John. Phasic policy gradient. In *International Conference on Machine Learning*, pp. 2020–2027. PMLR, 2021.

[9] Di Langosco, Lauro Langosco, Koch, Jack, Sharkey, Lee D, Pfau, Jacob, and Krueger, David. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

[10] Driess, Danny, Xia, Fei, Sajjadi, Mehdi SM, Lynch, Corey, Chowdhery, Aakanksha, Ichter, Brian, Wahid, Ayzaan, Tompson, Jonathan, Vuong, Quan, Yu, Tianhe, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[11] Gao, Peng, Geng, Shijie, Zhang, Renrui, Ma, Teli, Fang, Rongyao, Zhang, Yongfeng, Li, Hongsheng, and Qiao, Yu. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[12] Geng, Xinyang, Liu, Hao, Lee, Lisa, Schuurams, Dale, Levine, Sergey, and Abbeel, Pieter. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

[13] Huang, Wenlong, Abbeel, Pieter, Pathak, Deepak, and Mordatch, Igor. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022.

[14] Huang, Wenlong, Xia, Fei, Xiao, Ted, Chan, Harris, Liang, Jacky, Florence, Pete, Zeng, Andy, Tompson, Jonathan, Mordatch, Igor, Chebotar, Yevgen, Sermanet, Pierre, Jackson, Tomas, Brown, Noah, Luu, Linda, Levine, Sergey, Hausman, Karol, and brian ichter. Inner monologue: Embodied reasoning through planning with language models. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=3R3Pz5i0tye.

[15] Kim, Changyeon, Park, Jongjin, Shin, Jinwoo, Lee, Honglak, Abbeel, Pieter, and Lee, Kimin. Preference transformer: Modeling human preferences using transformers for RL. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Peot1SFDX0.

[16] Liu, Hao, Lee, Lisa, Lee, Kimin, and Abbeel, Pieter. Instruction-following agents with jointly pre-trained vision-language models. *arXiv preprint arXiv:2210.13431*, 2022.

[17] Liu, Yuqi, Xiong, Pengfei, Xu, Luhui, Cao, Shengming, and Jin, Qin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pp. 319–335. Springer, 2022.

[18] Loshchilov, Ilya and Hutter, Frank. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[19] Luo, Huaishao, Ji, Lei, Zhong, Ming, Chen, Yang, Lei, Wen, Duan, Nan, and Li, Tianrui. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[20] Pathak, Deepak, Gandhi, Dhiraj, and Gupta, Abhinav. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, 2019.

[21] Rasheed, Hanoona, khattak, Muhammad Uzair, Maaz, Muhammad, Khan, Salman, and Khan, Fahad Shahbaz. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[22] Seo, Younggyo, Chen, Lili, Shin, Jinwoo, Lee, Honglak, Abbeel, Pieter, and Lee, Kimin. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, pp. 9443–9454. PMLR, 2021.

[23] Shah, Dhruv, Osiński, Błażej, Levine, Sergey, 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.

[24] Wang, Yi, Li, Kunchang, Li, Yizhuo, He, Yinan, Huang, Bingkun, Zhao, Zhiyu, Zhang, Hongjie, Xu, Jilan, Liu, Yi, Wang, Zun, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.