Disentangled World Models: Learning to Transfer Semantic Knowledge from Distracting Videos for Reinforcement Learning

Supplementary Material

A. Compared Baselines

We compare DisWM with strong visual RL agents, including

- **DreamerV2** [10]: A model-based RL (MBRL) approach that trains world model and learns by imagining future latent states.
- **APV** [29]: It learns informational representations via action-free pretraining on videos and finetunes the agent with learned representations in the downstream tasks with action.
- **DV2 Finetune**: It pretrains a DreamerV2 agent [10] on distracting videos and then finetunes the trained model in the downstream tasks. Note that some tasks have different action spaces, which makes it difficult to finetune directly. Therefore, the action space of two tasks is set as the maximum action space of both environments.
- **TED** [7]: It adopts a classification task to learn temporally disentangled representations in visual RL.
- **CURL** [18]: A model-free RL method that employs contrastive learning to improve its sample efficiency.

B. Behavior Learning

For the behavior learning of DisWM, we adopt the actor-critic method following DreamerV2 [10]. Concretely, the actor and critic are both implemented as MLPs with ELU activations [5]. Formally, the actor and critic are defined as below:

$$\begin{aligned} &\text{Actor: } \hat{a}_t \sim \pi_\psi(\hat{a}_t|\hat{z}_t) \\ &\text{Critic: } v_\xi(\hat{z}_t) \approx \mathbb{E}_{p_\phi, p_\psi}\Big[\textstyle\sum_{\tau \geq t} \hat{\gamma}_{\tau-t} \hat{r}_\tau\Big]. \end{aligned} \tag{6}$$

The actor π_ψ is optimized by maximizing

$$\begin{aligned} \mathcal{L}(\psi) = \mathbb{E}_{p_\phi, p_\psi}\Big[\sum_{t=1}^{H-1} \big(&\underbrace{\beta \mathrm{H}\left[a_t \mid \hat{z}_t\right]}_{\text{entropy regularization}} + \underbrace{\rho V_t}_{\text{dynamics backprop}} \\ &+ \underbrace{(1-\rho) \ln \pi_\psi(\hat{a}_t \mid \hat{z}_t) \mathrm{sg}(V_t - v_\xi(\hat{z}_t))}_{\text{REINFORCE}} \big)\Big]. \end{aligned} \tag{7}$$

We train the critic v_ξ by minimizing

$$\mathcal{L}(\xi) = \mathbb{E}_{p_\phi, p_\psi}\Big[\sum_{t=1}^{H-1} \frac{1}{2}\left(v_\xi\left(\hat{z}_t\right) - \mathrm{sg}\left(V_t\right)\right)^2\Big]. \tag{8}$$

where sg is a stop gradient operator.

The λ-target V_t that involves a weighted average of reward information used in Eq. (7) and Eq. (8) is defined as:

$$V_t \doteq \hat{r}_t + \hat{\gamma}_t \begin{cases} (1-\lambda) v_\xi\left(\hat{z}_{t+1}\right) + \lambda V_{t+1} & \text{if } t < H \\ v_\xi\left(\hat{z}_H\right) & \text{if } t = H \end{cases}. \tag{9}$$

where H is the imagination horizon. Notably, the disentangled world model is *not* optimized during behavior learning.

C. Additional Results

C.1. Results on DMC

We compare the performance of *DreamerV3* [12], *TD-MPC2* [13], ContextWM [39], and our approach on DMC. As shown **Table A**, DisWM outperforms other strong baselines in terms of episode return.

C.2. Results on DrawerWorld

We present results on DrawerWorld [37] in **Table B**. As reported in **Table B**, DisWM (source: *Finger Spin*) outperforms other baselines in terms of success rate (%) on all tasks.

C.3. Sensitivity of the Latent Space Dimension

We visualize sensitivity analyses on the latent space dimension in Figure I. We observe that when $\mathbf{z}_{\text{dim}}$ for the β-VAE is too small, it impedes the learning of disentangled representations, leading to a decline in performance.

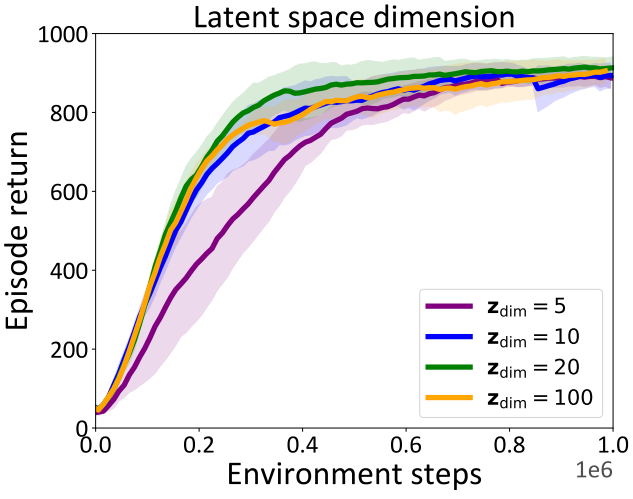

Figure I. Sensitivity analyses on *Cheetah Run* → *Walker Walk*

Model	*Reacher Easy* $\rightarrow$ *Cheetah Run*	*Walker Walk* $\rightarrow$ *Humanoid Walk*
DreamerV3	662 ± 9	12 ± 17
TD-MPC2	510 ± 15	1 ± 0
ContextWM	661 ± 49	1 ± 0
DisWM	$\mathbf{817 \pm 59}$	$\mathbf{147 \pm 85}$

Table A. Comparison with strong baselines on DMC.

Model	DrawerClose	DrawerOpen
TDMPC2	3 ± 6	43 ± 25
ContextWM	37 ± 12	23 ± 25
DisWM	$\mathbf{77 \pm 6}$	$\mathbf{70 \pm 10}$

Table B. Performance on DrawerWorld with texture variations.

C.4. Runtime Comparisons

We provide the detailed runtime and parameter comparisons with baselines in Table C. Note that the inference time is computed for one episode.

Table C. Runtime and model size comparisons evaluated on DMC (*Finger Spin* $\rightarrow$ *Reacher Easy*). DV2 FT is short for DreamerV2 finetune.

Model	Training Steps	Training time	Inference time	Params (M)
CURL	100k	303 min	4.97 sec	10.7
DV2 FT	200k	1522 min	9.88 sec	12.1
APV	200k	1722 min	10.15 sec	13
TED	100k	1051 min	20.49 sec	11.5
DV2	100k	901 min	9.59 sec	12.1
DisWM	200k	1311 min	9.48 sec	5.8

C.5. Sample Diversity Visualization

The adaptation stage enriches the sample diversity, as shown in Figure J, for *Cheetah Run* $\rightarrow$ *Walker Walk*, we sample 200 video clips of length 50 and visualize the corresponding latent features using t-SNE [25]. We find that the latent features of the online interactions are more diverse than those of the offline dataset.

D. Hyperparameters

The final hyperparameters of DisWM are reported in Table D.

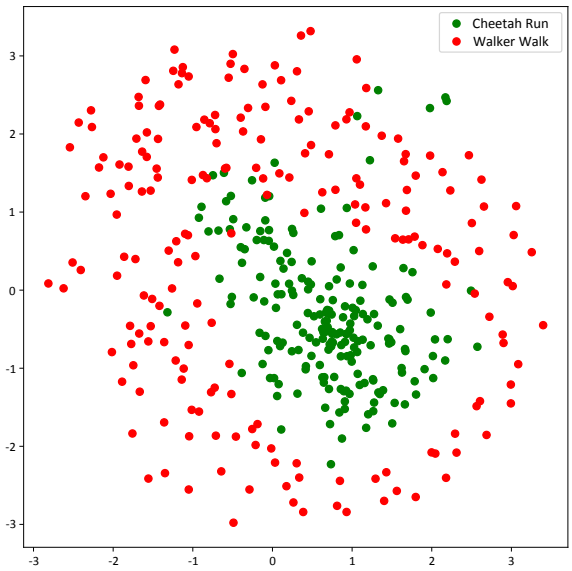

Figure J. Sample diversity enhanced by adaptation.

Table D. Hyperparameters of DisWM.

Name	**Notation**	**Value**
`Video prediction model`		
Image size	—	64×64
KL divergence scale	β_1	1
Disentanglement scale	β_2	0.015
Latent dimension	—	20
Learning rate	—	$3 \cdot 10^{-4}$
`Disentangled World Model`		
Latent distillation weight	η	0.1
Disentanglement scale	β	0.015
KL divergence scale	α	1
Latent dimension	—	20
Batch size	B	50
Batch length	L	50
Learning rate	—	$3 \cdot 10^{-4}$
`Behavior Learning`		
Imagination horizon	H	15
Discount	γ	0.99
λ-target	λ	0.95
Actor learning rate	—	$8 \cdot 10^{-5}$
Critic learning rate	—	$8 \cdot 10^{-5}$