

A IMPLEMENTATION DETAILS

We implement our algorithm and all baselines based on the codebase of C-bet (Cui et al., 2022). For WL-DM and V-BET, we consider only observations in the trajectory, and for V-DT and VIMA, we consider both observations and actions in the trajectory, which aligns with the implementation stated in the paper of C-bet (Cui et al., 2022), DT Chen et al. (2021b) and VIMA (Jiang et al., 2023). For WL-DM, V-BET, and V-DT, we use the same transformer model as stated in C-bet, which contains multiple self-attention layers to process video information and trajectory information at the same time. For VIMA, we use alternating cross-attention and self-attention layers as described in its paper (Jiang et al., 2023).

For all experiments, we set the learning rate to be 3×10^{-4} and set the window size for the trajectory to be 20 (for V-DT and VIMA, it means 20 observation-action pairs). For WL-DM, the window size of future video segments is sampled from $[20, 40]$. As we use the codebase of C-bet, all methods use the same action decoder, where we set the number of bins for action discretization to 32, and the id of each cluster will also be used for the representation of skills for WL-DM. For the Franka Kitchen environment (Gupta et al., 2020), we use decoders with 3 layers, and 3 heads and set the hidden dimension to be 60 (for VIMA, it means in total 3 self-attention layers and 3 cross-attention layers). We train all methods for 10 epochs. For WL-DM, α_1 is fixed to be 1×10^{-2} and α_2 is fixed to be 1×10^{-1} during the training process. For the Meta World environment (Yu et al., 2020), we use decoders with 6 layers, and 6 heads and set the hidden dimension to be 120 (for VIMA, it means in total 6 self-attention layers and 6 cross-attention layers). We train all methods for 30 epochs. For WL-DM, α_1 is set to be 0 in the beginning and fixed to be 1×10^{-3} after 10 epochs, and α_2 is fixed to be 10 during the training process.

B PROOF OF THEOREM 1

Theorem 1. *If we have $\text{MI}(h_v; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}) = 0$, then $D_{\text{KL}}(\pi(a|s, v) || \pi(a|s, \mathbf{v}_{\text{cur}})) = 0$ for all state-video pairs $(s, v) \in \mathcal{S} \times \mathcal{V}$ with non-zero probability $P(s, v) > 0$.*

Proof. By expanding the mutual information $\text{MI}(\mathbf{v}_{\text{other}}; h_v, a | s, \mathbf{v}_{\text{cur}})$, we can have the following equality:

$$\begin{aligned}
& \text{MI}(\mathbf{v}_{\text{other}}; h_v, a | s, \mathbf{v}_{\text{cur}}) \\
&= \mathbb{E}_{P(s, \mathbf{v}_{\text{cur}})} \mathbb{E}_{P(\mathbf{v}_{\text{other}}, h_v, a | s, \mathbf{v}_{\text{cur}})} \left[\log \frac{P(\mathbf{v}_{\text{other}}, h_v, a | s, \mathbf{v}_{\text{cur}})}{P(\mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}) P(h_v, a | s, \mathbf{v}_{\text{cur}})} \right] \\
&= \mathbb{E}_{P(s, \mathbf{v}_{\text{cur}})} \mathbb{E}_{P(\mathbf{v}_{\text{other}}, h_v, a | s, \mathbf{v}_{\text{cur}})} \left[\log P(h_v, a | s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}}) - \log P(h_v, a | s, \mathbf{v}_{\text{cur}}) \right] \\
&= \mathbb{E}_{P(s, \mathbf{v}_{\text{cur}})} \mathbb{E}_{P(\mathbf{v}_{\text{other}}, h_v, a | s, \mathbf{v}_{\text{cur}})} \left[\log P(h_v | s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}}) + P(a | h_v, s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}}) \right. \\
&\quad \left. - \log P(h_v | s, \mathbf{v}_{\text{cur}}) - \log P(a | h_v, s, \mathbf{v}_{\text{cur}}) \right] \\
&= \mathbb{E}_{P(s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}})} \left[D_{\text{KL}}(P(h_v | s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}}) || P(h_v | s, \mathbf{v}_{\text{cur}})) \right] \\
&\quad + \mathbb{E}_{P(h_v, s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}})} \left[D_{\text{KL}}(P(a | h_v, s, \mathbf{v}_{\text{cur}}, \mathbf{v}_{\text{other}}) || P(a | h_v, s, \mathbf{v}_{\text{cur}})) \right] \\
&= \text{MI}(h_v; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}) + \text{MI}(a; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}, h_v).
\end{aligned}$$

Similarly, we can also have:

$$\text{MI}(\mathbf{v}_{\text{other}}; h_v, a | s, \mathbf{v}_{\text{cur}}) = \text{MI}(a; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}) + \text{MI}(h_v; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}, a).$$

Combining these two equality, we can have:

$$\begin{aligned}
& \text{MI}(h_v; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}) + \text{MI}(a; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}, h_v) \\
&= \text{MI}(a; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}) + \text{MI}(h_v; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}, a).
\end{aligned}$$

As a and $\mathbf{v}_{\text{other}}$ become independent with each other when h_v is given, we have $\text{MI}(a; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}, h_v) = 0$. As we also have $\text{MI}(h_v; \mathbf{v}_{\text{other}} | s, \mathbf{v}_{\text{cur}}, a) \geq 0$, we can have the

following inequality, which basically gives us the conditional version of data processing inequality (Cover, 1999):

$$\text{MI}(h_v; v_{\text{other}} | s, v_{\text{cur}}) \geq \text{MI}(a; v_{\text{other}} | s, v_{\text{cur}}).$$

Since $\text{MI}(a; v_{\text{other}} | s, v_{\text{cur}}) \geq 0$, if we can also have $\text{MI}(h_v; v_{\text{other}} | s, v_{\text{cur}}) = 0$, then we can conclude that:

$$\text{MI}(a; v_{\text{other}} | s, v_{\text{cur}}) = 0.$$

By expanding this mutual information term, we have:

$$\begin{aligned} & \text{MI}(a; v_{\text{other}} | s, v_{\text{cur}}) \\ &= \mathbb{E}_{P(s, v_{\text{cur}}, v_{\text{other}})} \left[D_{\text{KL}}(\pi(a | s, v_{\text{cur}}, v_{\text{other}}) || \pi(a | s, v_{\text{cur}})) \right] \\ &= \mathbb{E}_{P(s, v)} \left[D_{\text{KL}}(\pi(a | s, v) || \pi(a | s, v_{\text{cur}})) \right] \\ &= 0. \end{aligned}$$

Since the KL divergence is non-negative, for the above expectation to be zero, there must be for all state-video pairs $(s, v) \in \mathcal{S} \times \mathcal{V}$ with non-zero probability $P(s, v) > 0$, we have the KL divergence to be zero, $D_{\text{KL}}(\pi(a | s, v) || \pi(a | s, v_{\text{cur}})) = 0$, and conclude our proof. \square

C VISUALIZATION

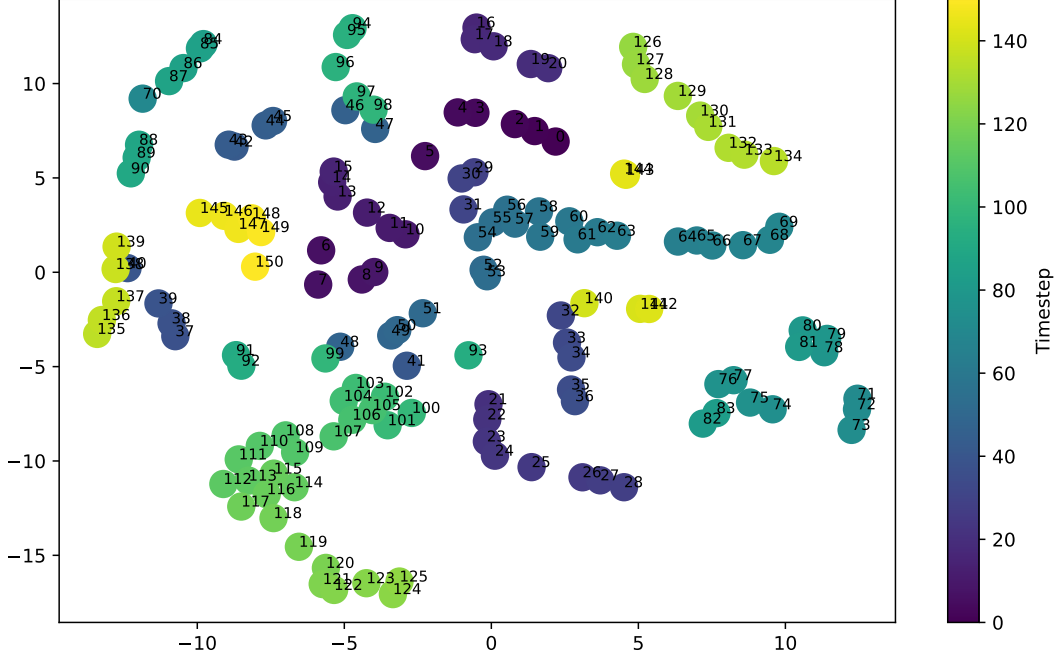


Figure 3: Visualization of h_v over timesteps.

In this section, we present the visualization result of our method. We visualize how h_v of WL-DL changes over timesteps. As shown in Figure 3, we can observe that h_v of WL-DM tends to converge at adjacent timesteps. It is worth noting that since we use a GPT-like transformer architecture as the encoder, the information of video tokens and obs tokens are mixed together in h_v . Furthermore, we do not introduce any task-level information (such as task-level video segmentation annotations), so the clustering results of h_v do not fully correspond to the task.

D ADDITIONAL EXPERIMENTS

D.1 ABLATION: TYPES OF TASK COMBINATIONS

We further include an experiment about the effect of the number of task combinations in the training set in the Meta World environment. In Section 5.2, we included 17 task combinations (7/3 split) in the training set, and here we further consider cases where we have 15 task combinations (6/4 split) and 20 task combinations in the training set. As shown Table 3, experimental results, WL-DM still outperforms other baselines, further demonstrating the effectiveness of our method.

	WL-DM	V-BET	V-DT	VIMA
6/4	1.93 ± 0.26	1.34 ± 0.78	0.86 ± 0.84	0.43 ± 0.78
7/3 (main exp)	2.57 ± 0.90	1.18 ± 0.85	1.24 ± 0.86	0.87 ± 0.84
8/2	1.88 ± 0.36	0.63 ± 0.86	0.94 ± 0.88	0.81 ± 0.61

Table 3: The performance of all methods with different number of task combinations in the training set on MW tasks.

D.2 ABLATION: NUMBER OF VIDEOS FOR EACH TASK COMBINATION

We also include an experiment about the number of videos corresponding to each task combination in the Meta World environment. In Section 5.2 we considered 20 different videos for each task combination, and here we further consider cases with 40 different videos for each task combination. As shown Table 4, experimental results, WL-DM still outperforms other baselines, which again demonstrates the effectiveness of WL-DM.

	WL-DM	V-BET	V-DT	VIMA
20 (main exp)	2.57 ± 0.90	1.18 ± 0.85	1.24 ± 0.86	0.87 ± 0.84
40	2.21 ± 0.81	1.71 ± 0.63	1.33 ± 0.90	1.09 ± 0.63

Table 4: The performance of all methods with different number of videos for each task combination on MW tasks.

D.3 MORE BASELINE: ViP

Env	Methods	Tasks							Avg
		ODWB	DOBW	DBWO	WBOD	BDOW	BDWO	BWDO	
MW	WL-DM	3.33	2.00	2.00	2.00	2.67	2.00	4.00	2.57 ± 0.90
	V-BET	1.87	2.00	0.73	1.33	0.33	0.00	1.97	1.18 ± 0.85
	V-DT	1.33	2.13	1.23	1.93	0.37	0.83	0.83	1.24 ± 0.86
	VIMA	1.80	1.00	0.37	0.37	1.17	0.57	0.83	0.87 ± 0.84
	ViP	2.80	2.20	2.00	1.87	1.63	1.10	1.80	1.91 ± 0.88

Table 5: The performance of all methods including ViP on all MW tasks.

We have added a new video-conditioned baseline: ViP (Chane-Sane et al., 2023). Although the purpose of ViP is to learn a video-conditioned policy, it has a different setting from our method. Thus, we have made the following modifications to adapt it to our setting:

- Since we do not consider human videos as input, we have removed the part that uses human videos for pre-training.

- Since we do not assume access to video labels, we have changed its supervised contrastive learning part to unsupervised contrastive learning on robot videos.

We used the same codebase as WL-DM to implement ViP with minimal modifications, and conducted experiments in the MetaWorld environment. The experimental results are shown in Table 5. ViP outperformed other baselines in this environment, demonstrating its effectiveness as video-conditioned policy. However, its performance still lags behind WL-DM, which further demonstrates the effectiveness of WL-DM.

D.4 MORE DATASET

To further evaluate our method, we construct script in a similar way of Lee et al. (2024) to convert state-base observations of the original dataset of the Franka Kitchen environment into image-based observation, and train all methods on this dataset. For WL-DM, we use a linear schedule for α_1 , where coef_start is set to be 0 and coef_end is set to be 1×10^{-4} , and for α_2 , we fix it to be 1×10^{-1} during the training process. As shown in Table 6, WL-DM still outperforms other methods, which further demonstrates the effectiveness of our method.

Env	Methods	Tasks							Avg
		BTLS	BTSH	MBTS	MBTH	MLSH	MBTL	MKBH	
FK(new)	WL-DM	2.43	1.63	2.63	2.57	2.27	2.27	2.5	2.33 ± 0.74
	V-BET	1.23	1.43	2.33	1.53	2.10	1.70	2.17	1.79 ± 0.80
	V-DT	1.63	2.20	1.40	1.80	1.47	1.73	2.20	1.78 ± 0.83
	VIMA	0.87	1.27	2.20	1.80	1.80	1.43	2.23	1.66 ± 0.80

Table 6: The performance of all methods on new FK dataset.

D.5 MORE BASE ALGORITHM

As WL-DM can be seen as a method using information bottleneck-based loss on top of V-BET. To further validate our approach, we applied the information bottleneck-based loss of WL-DM to both V-DT and VIMA and conducted experiments in the Meta World environment. As shown in Table 7, WL-DM+V-DT and WL-DM+VIMA both outperform its base algorithm, which further validates the effectiveness of the proposed information bottleneck-based loss.

	WL-DM	V-BET	WL-DM+V-DT	V-DT	WL-DM+VIMA	VIMA
MW	2.57 ± 0.90	1.18 ± 0.85	1.72 ± 0.68	1.24 ± 0.86	1.84 ± 0.84	0.87 ± 0.84

Table 7: The performance of all methods on MW tasks, we applied the information bottleneck-based loss on different base algorithms and compare their performance.

D.6 COEFFICIENT SELECTION

$\alpha_1 \backslash \alpha_2$	0.1	1	10
0.1	0.95 ± 0.77	2.09 ± 0.69	2.16 ± 0.71
0.01	0.82 ± 0.86	1.83 ± 0.53	2.05 ± 0.62
0.001	1.03 ± 0.80	1.79 ± 0.41	2.57 ± 0.90

Table 8: The performance of WL-DM with different coefficients on MW tasks.

Coefficients for mutual information loss should be adjusted according to the environment. Specifically, we conducted a grid search to select optimal coefficients. Taking experiments in Meta World environment as an example, the performance of different coefficients is shown in Table 8.