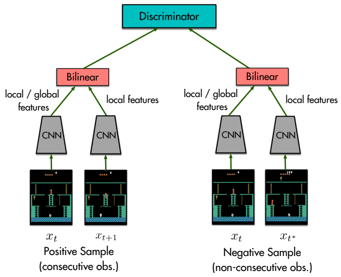
We thank all the reviewers for their thoughtful responses. As **R1** and **R3** note, we tried to focus on the benchmark and evaluations within it. **R2** highlighted important questions about some design choices in the benchmark; we have made active improvements on some of them, and better explained our rationale on others below. We have also expanded on our discussion of ST-DIM in the draft, and included an additional figure to explain the contrastive task. We address below some of the concerns that were raised.



ST-DIM vs DIM (R1): DIM was originally introduced in the context of static images. In this work, we extend DIM to work with temporal data by contrasting local and global features across frames at different time steps instead of within the same image.

Real-World Setups (R1): Evaluating how well these algorithms transfer to the real-world is an exciting direction. Since contrastive methods don't focus on pixel-level details, we expect them to transfer better than generative methods. Moreover, practical applications such as robotics are inherently spatio-temporal, which make them conducive for methods like ST-DIM that exploit such structure.

Coverage in AARI (R2): For multi-game rooms (MZR, Venture, Pitfall, Private-Eye, Hero), we have expanded our RAM labels to include objects and enemies from all the rooms. For example, RAM indices 44, 45 in MZR now correspond to: the key (in rooms 1,7,8,14), spider (rooms 4,13,21), torch (5), sword (6), snake (9,11,22), jewel (0,10,15,23), amulet (19), bouncing skulls (2,3). These labels are now grouped by room, which allows us to focus only on variables relevant in a room, and ignore other spurious/null values.

Data Collection (R2): Our default data collection mode now also includes a PPO agent trained for 50M steps (results below) over full observations, which significantly improves coverage over the 10M agent. We want to stress that the default data collection modes are meant only for evaluating different representation algorithms, and we think 22 games over multiple rooms provide adequate visual diversity to do so. Having said that, our RAM interface does expose labels from all the rooms now, so exploration methods or new methods can still leverage them to make systematic evaluations.

	MAJ-CLF	RANDOM-CNN	VAE	PIXEL-PRED	CPC	ST-DIM	SUPERVISED
MEAN	0.11	0.38	0.38	0.54	0.56	**0.61**	0.80

N/A Results(R2): The missing values in Table 2 for supervised are: 0.60 for Asteroids, and 0.85 for Boxing.

Related papers (R2): Thanks for pointing us to the relevant papers. We have now included them in Section 2 and they have significantly improved our discussion of related work.

Evaluation Metrics (R2): We want to stress that the output labels in our dataset are ordered, but not continuous (int8's). Our use of cross-entropy for ordinal classification follows a variety of prior work that have used CE in lieu of MSE regression [Colorful Colorization (Zhang 2016), DIM (Hjelm 2018, sec 4.3) etc.]. We use F1 score over accuracy because the dataset has uneven label distribution (e.g. in Pitfall the class, "player y position" is more frequently 32 (ground level) than 21 (peak of a jump)). We have also added accuracy numbers to the appendix (summary results below). Specifically, we use F1 score with average='weighted' from sklearn, this averages F1 score from each class weighted by support. We will also add class distribution numbers (min, max, mode, etc.).

	MAJ-CLF	RANDOM-CNN	VAE	PIXEL-PRED	CPC	ST-DIM	SUPERVISED
MEAN	0.24	0.48	0.43	0.60	0.62	**0.69**	0.82

Non-temporal DIM baseline (R2): We ablated the temporal aspect in ST-DIM by removing any temporal cues from the contrastive task, making the setup closer to DIM. We found it to perform much worse than ST-DIM indicating that temporal cues are really important. We will add a game-by-game plot similar to other ablations in our updated draft.

Open-source (R3): Open-sourcing our code and benchmark is a top priority for us. In fact, since the submission, cleaning up the code and improving the API have been our key focus. We are excited to make a public release soon!

Interpretation of Results (R3): Folk wisdom does say that contrastive methods should do better at small objects (than generative models); however, we couldn't find such claims empirically verified. Moreover, we believe discussing robustness to easy-to-exploit features is important as contrastive methods become more popular. Holistically, we systematically compare contrastive and generative methods, insights from which should be helpful to the RL community which has mostly focused on using generative methods such as VAEs for representation learning.

ST-DIM's empirical superiority to CPC (R3): ST-DIM performs better than CPC because it avoids the trap of focusing on a single easily predictable factor [WDM (Ozair et. al, 2019)]. In Table 4, we see that both Global-T-DIM and CPC focus only on easy factors like the clock and the score, but ST-DIM captures other factors as well, indicating local objectives (the only difference b/w Global-T-DIM and ST-DIM) help make it robust to easy-to-exploit factors.