Abstract—Temporal event representations are an essential aspect of learning among humans. They allow for succinct encoding of the experiences we have through a variety of sensory inputs. Also, they are believed to be arranged hierarchically, allowing for an efficient representation of complex long-horizon experiences. Additionally, these representations are acquired in a self-supervised manner. Analogously, here we propose a model that learns temporal representations from long-horizon visual demonstration data and associated textual descriptions, without explicit temporal supervision. Our method produces a hierarchy of representations that align more closely with ground-truth human-annotated events (+15.3%) than state-of-the-art unsupervised baselines. Our results are comparable to heavily-supervised baselines in complex visual domains such as Chess Openings, YouCook2 and TutorialVQA datasets. Finally, we perform ablation studies illustrating the robustness of our approach. We release our code and demo visualizations in the Supplementary Material.

I. INTRODUCTION AND RELATED WORK

Multi-modal Event Representation learning in demonstration videos is a challenging problem. Existing methods attempt to chunk videos into events using prohibitively expensive, heavily annotated datasets containing labels for objects per frame and timestamps for activities in the video. Further, existing methods suffer from sub-optimal performance when learning event representations for long sequences. In contrast, humans excel in such scenarios - given a video demonstration (Figure 1) of a complex task (such as cooking), humans can subconsciously abstract events (such as boiling, frying, pouring, etc.) that succinctly encode sub-sequences in such demonstrations (1). These events are hierarchical in nature - lower-level events are building blocks for higher-level events (2).

To learn such hierarchical events, we propose an end-to-end trainable Seq2Seq architecture, SHERLock (Self-supervised Hierarchical Event Representation Learning), for multi-modal hierarchical representation learning from demonstrations. SHERLock takes a long-horizon sequence of demonstration images (in our case, chess, tutorial, and cooking) and commentary as input. It can then isolate semantically meaningful subsequences in input trajectories. Through ablations, we show how variants of SHERLock discover meaningful subsequences using only a sequence of images. Our method does not require timestamps of video and commentary, nor does it need any alignment annotation between video and textual inputs. We only assume the order of the events and narration are preserved in the input data. Our architecture discovers event representations along with their hierarchical organization without any supervision.

SHERLock improves upon the state-of-the-art of related work in the following ways:

1) Self-supervised: State-of-the-art approaches in allied fields (3; 4; 5; 6; 7; 8) (skill learning, event detection etc.) require large datasets of demonstrations, with expensive human annotations for timestamps corresponding to each event. (9) discover motor primitives from demonstrations but in a non-hierarchical fashion. SHERLock, on the other hand, abstracts hierarchical event representations from multimodal data, i.e. it divides long-horizon trajectories into a hierarchy of semantically meaningful subsequences, without requiring any temporal annotations.

2) Long Horizon: Long-horizon tasks remain the bane of learning systems, due to an aggregation of sub-optimal behavior over a horizon (10). Previous works in imitation learning (11; 12; 13; 14; 15; 16), show how agents can learn representations for events in simple tasks like cart-pole from demonstrations. More recently, (17) shows that agents can learn action representation using a large corpus of observation data, i.e., trajectories of states and a relatively smaller corpus of interaction data, i.e., trajectories of state-action pairs. However, these approaches all restrict themselves to short horizons, while SHERLock is able to generate meaningful event representations for long-horizon tasks like cooking and chess.
3) **Offline abstraction:** Recent works in unsupervised skill/event discovery (18; 19; 20; 21) require costly interconnected actions with an environment to discover skill sequences, an infeasible assumption in domains such as cooking or healthcare, where exploration is potentially dangerous. (18) learn a large number of low-level sequences of actions by forcing the agent to produce skills that are different from those previously acquired. Similarly, (19) attempt to learn skills such that their transitions are almost deterministic in a given environment. However, these approaches require access to an environment while SHERLock discovers these representations from offline demonstrations, utilizing large amounts of demonstration data.

4) **View Invariance:** SHERLock abstracts events from demonstrations of a variety of cooking tasks. The demonstration videos originate from a number of sources, varying in camera-angles, instructional styles, etc. Recent works in unsupervised skill/event learning are more restrictive (9; 18; 19), requiring that demonstration data originate from a single viewpoint, with coincident state and action spaces.

5) **Multi-modality and Interpretability:** SHERLock learns a joint latent space for events utilizing both textual and visual inputs which are available in typical human demonstrations. This allows us to both visualize the physical manifestation of a temporal event and describe in words the outcome. This is an improvement upon recent works in unsupervised skill learning, which utilize demonstrations corresponding to low dimensional state spaces and simple control signals (16; 14; 20; 21).

6) **Hierarchical:** We find that hierarchical events abstracted by SHERLock are indeed more semantically meaningful and align more closely with ground-truth annotations for events in real-world datasets (YouCook2 (22), Chess Opening and TutorialVQA (23)) than other non-hierarchical approaches (9). See **Non-Hierarchy** in Table I.

## II. APPROACH

### A. Overview

We explain the motivation for SHERLock with an example in the domain of cooking demonstrations. Consider a long horizon demonstration for example, of an Eggs Benedict recipe. Here, low-level events might include boiling water or addition of eggs to water. Several such low-level events may combine to produce a high-level event - e.g., poaching an egg, which consists of boiling water, addition of egg to water, and finally removal after two minutes of cooking. SHERLock learns embeddings for such low and high-level events.

Broadly, SHERLock (Figure 2) can be described as a multi-modal, hierarchical, sequence-to-sequence model. The model receives as input a sequence of pre-trained ResNet-50 Embeddings ((24)), in addition to a sequence of pre-trained BERT-base (25) embeddings. The two modalities are encoded separately by two transformer models into a pair of sequences of low-level latent event embeddings (e.g., boiling water or placing eggs in water, derived from either video or text). Such low-level sequences are further encoded by another pair of transformers that generate sequences of high-level event embeddings (e.g., poaching an egg). The embedding pairs are aligned through an L2 loss, forcing both representations to correspond to one another. Subsequently, a cross-modal decoding scheme is implemented: visual embeddings are used to re-generate word / BERT-base embeddings, while textual embeddings are used to generate video frame ResNet embeddings. After successful training, the system hence is expected to generate modality and domain invariant embeddings for temporal events. Those embeddings could subsequently be used for event classification and robotic skill learning (to be developed in future work).

### B. Hierarchical Events

Intuitively, we define an event as a short sequence of states which may occur repeatedly across several demonstration trajectories. Events have an upper limit on their length in time steps. They can be obtained from both a sequence of demonstration images (S = s0:mn) and from the associated textual description (W = w0:n). Additionally, they are hierarchical in nature - thus, low-level and high-level events representations are denoted by zL and zH, respectively (while the following discussion is restricted to two levels, we explore the effect of more levels in Table II). Given a low-level event representation, an associated sequence (of words or images) can be obtained using a decoder \( \Phi^{L-dec} \):

\[
x_t | z_L^t \sim N(\mu_{L,t}, \sigma_{L,t}^2)
\]

where \( [\mu_{L,t}, \sigma_{L,t}^2] = \Phi^{L-dec}(z_L^t, x_{t-1}) \) (1)

where \( X = x_{0:T} \) may correspond to the flattened embedding of words \( W \) or images \( S \), and \( N(\cdot|\cdot) \) is a Gaussian distribution (assume prior) with parameters generated by the neural network \( \Phi^{H-dec} \). Events also exhibit a temporal hierarchy. High-level events are generated as:

\[
z_H^t | z_L^{t_1} \sim N(\mu_{H,t}, \sigma_{H,t}^2)
\]

where \( [\mu_{H,t}, \sigma_{H,t}^2] = \Phi^{H-dec}(z_H^{t_1}, z_L^{t_1}) \) (2)

Given such a high-level event \( z_H^t \), the associated sequence of low-level events can be approximated through a function \( \Phi^{L-dec} \) as:

\[
z_L^t | z_H^t, z_L^{t_1} \sim N(\mu_{L,t}, \sigma_{L,t}^2)
\]

where \( [\mu_{L,t}, \sigma_{L,t}^2] = \Phi^{L-dec}(z_H^t, z_L^{t_1}) \) (3)

Thus, the resulting joint model mapped over trajectories of images \( p(S, z_L^t, z_H^t) \) factorizes as:

\[
p(s_0) \prod_{t=1}^{mn} p(s_t | z_L^{t_1}, s_{t-1}) p(z_L^t | z_L^{t_1}, z_H^t) p(z_H^t | z_L^{t_1})
\]

(4)

and the resulting joint model mapped over trajectories of words \( p(W, z_L^t, z_H^t) \) factorizes as:

\[
p(w_0) \prod_{t=1}^{n} p(w_t | z_L^{t_1}, w_{t-1}) p(z_L^t | z_L^{t_1}, z_H^t) p(z_H^t | z_L^{t_1})
\]

(5)
The transition functions \( p(z^L_x|z^H_x, z^L_y) \) and \( p(z^H_y|x^0) \) are also learned using fixed length transformer models (26).

C. Architecture

SHERLock consists of 2 pairs of encoding transformers - one pair for each of the modalities. For a modality \( X \), where \( X \in \{ \text{images, words} \} \), the pair of encoders consists of \( q(z^L_x|X) \), which encodes the modality \( X \) into low-level events \( z^L_x \) and \( q(z^L_y|L) \) which encodes low-level events \( z^L_y \) into high level events \( z^H_y \).

\[
z^L_x = q(z^L_x|X) \quad \text{and} \quad z^H_y = q(z^H_y|x^0)
\]

(6)

Analogously, SHERLock also contains 2 pairs of decoding transformers - one pair for each of the modalities. Decoding occurs in a cross-modal manner - textual events generate video and visual events generate text. Thus, for a modality \( X \), where \( X \in \{ \text{images, words} \} \), \( p(z^L_x|x^0) \) generates low-level supervision from high-level events of the modality \( X \) and \( p(x'|z^L_y) \) regenerates the modality \( X \).

\[
z^L_x = p(z^L_x|x^0, z^H_x) \quad \text{and} \quad x' = p(x'|z^L_y)
\]

(7)

D. Training Metrics

1) Soft-Dynamic Time Warping (Soft-DTW): Given two trajectories \( x = (x_1, x_2, \ldots, x_N) \) and \( y = (y_1, y_2, \ldots, y_M) \), the soft-DTW \( (x, y) \) (27)) computes the discrepancy between \( x \) and \( y \) as

\[
\text{soft-DTW}(x, y) = \min_{A, \Delta(x, y)} \{ \langle A, \Delta(x, y) \rangle : A \in A_{n,m} \}
\]

(8)

where \( A \in A_{n,m} \) is the alignment matrix, \( \Delta(x, y) = [\delta(x_i, y_j)]_{i,j} \in \mathbb{R}^{n \times m} \) and \( \delta \) being the cost function. \( \min_{A, \Delta} \) operator is then computed as,

\[
\min_{A, \Delta} \{ \min_{1 \leq n \leq m} \{ a_i, \ldots, a_n \} \} = \begin{cases} \min_{1 \leq n \leq m} a_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma}, & \gamma > 0. \end{cases}
\]

(9)

For our experiments, we use \( L_2 \) distance as \( \delta \) and \( \gamma = 1 \).

2) Learning Objective: We emphasize that we do not require supervision for hierarchical temporal segmentation, i.e., we do not require annotations which demarcate the beginning and ending of an event, both in language and in the space of frame’s timestamps. Our approach uses several loss terms between network outputs to achieve our objective.

\[
L_{\text{dyn}} = \text{soft-DTW}(Z^L_x, Z^H_w) + \text{soft-DTW}(Z^L_x, Z^L_y) + \text{soft-DTW}(W, W') + \text{soft-DTW}(Z^H_x, Z^H_w) + \text{soft-DTW}(Z^L_y, Z^L_x)
\]

(10)

We then define our total loss as, \( L_{\text{total}} = L_{\text{dyn}} + \beta \cdot L_{\text{static}} \). We posit that this loss function provides the inductive bias necessary for learning the event latent space. The term soft-DTW \( (S, S') \) ensures reconstruction of demonstration frames from the textual events, while soft-DTW \( (W, W') \) ensures the generation of textual description from visual events.

\[
L_{\text{static}} = L_2(Z^H_x, Z^H_w) + L_2(Z^L_x, Z^L_y)
\]

(11)

E. Evaluation Metrics

The ground-truth events in the dataset and the events generated by SHERLock may differ in number, duration, and start-time. To evaluate the efficacy of SHERLock in generating events that align with the human-annotated events in our dataset, it is imperative that we utilize a metric that measures the overlap between generated events and ground truths and also accounts for this possible temporal mismatch.

Consider the search series \( X = (x_1, x_2, \ldots, x_M) \) and target series \( T = (t_1, t_2, \ldots, t_N) \) where \( X \) corresponds to the end-of-event time stamp for each event as generated by SHERLock for a single long-horizon demonstration trajectory. Thus, the \( i^{th} \) event abstracted from SHERLock starts at time \( x_{i-1} \) and end at time \( x_i \). Similarly, \( T \) corresponds to the end-of-event time stamp for each ground-truth event in the demonstration trajectory, where the \( j^{th} \) ground truth event starts at time \( t_{j-1} \) and ends at time \( t_j \). Note that both \( x_0 \) and \( t_0 \) are equal to...
Fig. 3. t-SNE of low-level events and their corresponding high-level mappings discovered by SHERLock on the YouCook2 dataset. We obtain clusters of low-level events such as frying, pouring while frying, seasoning etc. We also obtain two high-level events that correspond to events that require heating and those that do not.

zero i.e. we measure time starting at zero for all demonstration trajectories.

To meaningfully compute the intersection over union (IoU) between ground truth and outputs from SHERLock, we first need to align the two representations using dynamic time warping (DTW; (28)). This implies calculating $\Delta(X, T)$, solving the following DTW optimization problem $((28))$, $\Delta(X, T) = \min_{P} \sum_{m \in P} \delta(x_m, t_n)$.

where the $X$ and $T$ correspond to the search and target series respectively and $\delta$ corresponds to a distance metric (in our case the $L_2$ norm), measuring time mismatch.

$\Delta(X, T)$ therefore corresponds to the trajectory discrepancy measure defined as the matching cost for the optimal matching path $P$ among all possible valid matching paths $P$ (i.e., paths satisfying monotonicity, continuity, and boundary conditions). From this optimal trajectory we can also obtain the warping function $W$ such that $W(x_i) = t_j$, i.e. we find the optimal mapping between the $i^{th}$ event ending at time $x_i$ and the $j^{th}$ event ending at time $t_j$. The resulting Intersection over Union for a single long-horizon trajectory, Time-warped IoU (TW-IoU), is:

$$\sum_{t_i} \sum_{x_j} W(x_j) = t_i \min(t_i, x_j) - \max(t_{i-1}, x_j-1)$$

More details are presented in Section D of Supplementary Material.

F. Alignment during Inference

We calculate the DTW path (28) ($\gamma = 0$ case in eqn. (9)) between a decoded sequence and a ground truth video to obtain the optimal alignment between ground truth video frames and predicted video frames (high & low-level). This alignment is subsequently used during the calculation of the TW-IoU scores.

III. EXPERIMENTS

Datasets: YouCook2((22)) dataset comprises of instructional videos for 89 unique food recipes. Recommending Chess Openings1 dataset consists of opening moves in the game of Chess. TutorialVQA ((23)) consists of 76 tutorial videos pertaining to an image editing software. For dataset details, see Sec B in the Appendix. For implementation details, see Sec A in the Appendix.

A. Visualizing Hierarchy

Here we analyze whether the discovered events are human interpretable i.e. are the temporal clusters within a single demonstration semantically meaningful? We find that SHERLock abstracts several useful human interpretable events.
### Ablation Variants TW-IoU

| Method | TW-IoU |
|--------|--------|
| SHERLock w/o comment w/o L2 loss | 38.46 |
| SHERLock w/o comment w/o cross-decoding | 37.67 |
| SHERLock Single-level Decoding | 20.33 |
| SHERLock w/o comment w/o low-align loss | 18.99 |
| SHERLock Three-level Hierarchy | 37.61 |
| SHERLock w/o comment (200 Frames) | 39.45 |
| SHERLock w/o comment (64 Frames) | 33.41 |
| SHERLock w/o comment (32 Frames) | 12.79 |

**Table II**

TW-IoU scores for ablation experiments. Note that the reported TW-IoU scores are calculated with reference to high-level annotations available in the dataset (low-level annotations are unavailable). See Figure 8 of Supplementary Material for model architectures.

### Comparison with Baselines

We evaluate the performance of SHERLock quantitatively on YouCook2 and TutorialVQA and quantify its ability to generate coherent events that align with the human annotated ground truths using the TW-IoU metric. We compare our approach with 6 baselines.

**GRU Time Stamp Prediction:** A supervised baseline comprising of a GRU-based encoder (31) that sequentially processes the ResNet features corresponding to frames in a video followed by a decoder GRU (32) that attends to encoder outputs and is trained to sequentially predict end-of-event timestamps of each meaningful segment (variable in number) in the video.

**Non-Hierarchical w/ comment:** We implement the (9) approach (SOTA in unsupervised skill learning w/o environment) which takes as input a sequence of video frames and discovers a single level of events without any hierarchy.

**Non-Hierarchical w/o comment:** A modified multi-modal version of Non-Hierarchical where frames and words are utilized to form a non-hierarchical latent event representation. This baseline ascertains the effect of both hierarchical and multi-modal learning on the representations obtained. See Fig 7 in Supplementary Material for architectural details.

**Clustering - ResNet32 Embeddings:** Given an input sequence of frames, we define the weight function based on their temporal position in the sequence and also the $L_2$ distance between the frame embeddings. Then we use standard K-means algorithm (we find best K=4) to cluster the frames based on the weighting function defined and use the clusters formed to predict the temporal boundaries.

**Clustering - HowTo100M Embeddings:** We utilize the pre-trained embeddings from the supervised action recognition dataset and method (29) and apply a K-means (we find best K=4) clustering on them.

**GRU Supervised Segment Prediction:** Instead of predicting end time stamps of each segment (as in GRU Time Stamp Prediction), the decoder is trained to predict/assign identical ids to frames which are part of the same segment. Further, the model’s decoder is trained to assign different ids to frames part of different segments while frames not part of any meaningful segment in the ground truth are trained to have a default null id - 0.

Table I summarises and compares the TW-IoU computed between ground truth time stamp annotations and predicted/discovered segments. SHERLock achieves the highest TW-IoU when compared with all other unsupervised baselines. We find that SHERLock discovers events that align better with the ground truth events (SHERLock performs ~ 23% better) compared to Non-Hierarchical (9) performing at par with the supervised baselines.

### Ablation Experiments

**Effect of Sampling Rate on the quality of hierarchy** For YouCook2, we cap the length of a frame sequence to 200 frames (down-sampled from the original frames provided in the dataset due to memory constraints). Subsequently, we analyze the trade-off between sequence length and performance. This provides an insight into granularity of information required to discover naturalistic hierarchies. Interestingly, we don’t observe a linear drop in performance with a reduction in the number of frames (refer Table II).

**Effect of Guidance through Commentary** We study the
effect that language has on event discovery by comparing SHERLock without comment (Fig 6 in Supplementary), which discovers event hierarchy using just frames and SHERLock which additionally uses word embeddings as a guide (as in Figure 2). Language improves the TW-IoU by $\sim 10\%$, indicating that using commentary enables SHERLock to detect more precisely the boundaries of segments corresponding to various events in a trajectory. Further, we find (Fig-7 in supplementary) that the implicitly hierarchical nature of the language provides inductive bias to the model to learn a more natural hierarchy of events.

**Number of Levels in Hierarchy** We explore the effect of a third level of hierarchy, through additional transformers during the encoding and decoding phase. Thus, our architecture generates 16 low-level, 8 mid-level and 4 high level events. We find that this third level of event provides only a marginal improvement over the TW-IoU scores which we report in Table II. Additionally, we find that this increases the GPU memory requirements during training due to the increased number of model parameters in memory along with the additional losses (calculating Soft-DTW losses means solving a dynamic programming problem).

**Model Complexity:** SHERLock uses the Transformer architecture for modeling $\Phi$ and $p()$ which makes the model a bit heavy to train. So, we experiment by replacing all the transformer modules with simpler GRU modules keeping same number of layers (See SHERLock-GRU in Table I). We observe that there is not much difference in performance ($\sim 3.5\%$). Also it still outperforms all other unsupervised baselines. This indicates that the attention mechanism in Transformers does help us learn better representations but most of the gain can be attributed to the model architecture.

**Components and Losses:** We perform ablation experiments to ascertain the need for each of the modules and losses used in SHERLock. We remove the soft-DTW($Z_L^L$, $Z_L^L$) loss from our SHERLock to highlight its importance in maintaining the fidelity of the reconstruction scheme. This loss guides the alignment between the encoded low-level events ($Z_L^L$) and the reconstructed low-level events ($Z_L^L$). We find that removing this loss reduces the TW-IoU scores drastically (see SHERLock w/o comment w/o low-align loss in Table II).

We also evaluate a simplified version of the SHERLock w/o commentary model, where we remove the $Z_L^L = z_{0:T}^L \sim p(z_L^L|z_H^H)$ modules and re-generate the word and visual sequence embeddings from the high-level events as $X' = x_{0:T}^{X'} \sim p(x'|z_H^H)$. We see this results in the drop of TW-IoU (Table II), thus confirming our need for the step-wise encoding-and-decoding scheme used. We call this the SHERLock Single-level Decoding baseline the diagram for which is included in the Supplementary Material (Fig-7).

**IV. CONCLUSION**

In this paper, we provide a self-supervised method (SHERLock) capable of hierarchical and multi-modal learning. It can discover events and organize them in a meaningful hierarchy using only demonstration data from chess openings, tutorials, and cooking. We also show that this discovered hierarchy of events helps predict textual labels and temporal event segmentations for the associated demonstrations.

One limitation is that we can’t use longer video sequences for training since computing soft-DTW requires solving a DP problem in quadratic space. Also sometimes specific nouns like "lobster" are replaced with more commonly appearing nouns like "patty", which is due to the fact that grounding of nouns using a few images is very difficult. This could be an interesting direction for future work. Also, we would explore curriculum learning where the discovered event hierarchy by SHERLock is used to generate curricula where lower-level events would be taught first followed by higher-level events and also be used for option discovery and training in reinforcement learning.

**REFERENCES**

[1] K. B. R. S. Pammi, Chandrasekhar and K. Doya, “Chunking phenomenon in complex sequential skill learning in humans,” in International Conference on Neural Information Processing. Springer, 2004, pp. 294–299.
[2] M. S. Naim, Michelangelo and M. Tsodyks, “Emergence of hierarchical organization in memory for random material,” Scientific Reports, vol. 9, no. 1, pp. 1–10, 2019.

[3] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, “Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation,” in CVPR, 2019, pp. 2624–2632.

[4] A. W. Boggust, K. Audhkhasi, D. Joshi, D. Harwath, S. Thomas, R. S. Feis, D. Gutfreund, Y. Zhang, A. Torralba, M. Picheny et al., “Grounding spoken words in unlabeled video,” in CVPR Workshops, 2019, pp. 29–32.

[5] F. Tosi, F. Aleotti, P. Z. Ramirez, M. Poggi, S. Salti, L. D. Stefano, and S. Mattoccia, “Distilled semantics for comprehensive scene understanding from videos,” in CVPR, 2020, pp. 4654–4665.

[6] A. Miech, J. B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, “End-to-end learning of visual representations from uncurated instructional videos,” CVPR, 2020.

[7] L. Zhu and Y. Yang, “Actbert: Learning global-local video-text representations,” CVPR, 2020.

[8] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “Videobert: A joint model for video and language representation learning,” ICCV, 2019.

[9] T. Shankar, S. Tulsiani, L. Pinto, and A. Gupta, “Discovering motor programs by recomposing demonstrations,” in ICLR, 2020. [Online]. Available: https://openreview.net/forum?id=rkgHY0NYwr

[10] M. N. Nicolescu and M. J. Mataric, “Natural methods for robot task learning: Instructive demonstrations, generalization and practice,” in AAMAS, 2003, pp. 241–248.

[11] N. Schaal, “Learning from demonstration,” in Advances in neural information processing systems, 1997, pp. 1040–1046.

[12] N. Esmaili, C. Sammut, and G. Shirazi, “Behavioural cloning in control of a dynamic system,” in 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, vol. 3. IEEE, 1995, pp. 2904–2909.

[13] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, “Learning and generalization of motor skills by learning from demonstration,” in ICRA. IEEE, 2009, pp. 763–768.

[14] S. Niekum, S. Osentoski, G. Konidaris, and A. G. Barto, “Learning and generalization of complex tasks from unstructured demonstrations,” in IROS, 2012.

[15] C. G. Atkeson and S. Schaal, “Robot learning from demonstration,” in ICML, vol. 97. Citeseer, 1997, pp. 12–20.

[16] J. Peters, J. Koerber, K. Mulling, O. Kramer, and G. Neumann, “Towards robot skill learning: From simple skills to table tennis,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2013, pp. 627–631.

[17] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Danilidis, S. Levine, and C. Finn, “Learning predictive models from observation and interaction,” arXiv preprint arXiv:1912.12773, 2019.

[18] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, “Diversity is all you need: Learning skills without a reward function,” arXiv preprint arXiv:1802.06070, 2018.

[19] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, “Dynamics-aware unsupervised discovery of skills,” arXiv preprint arXiv:1907.01657, 2019.

[20] D. Xu, S. Nair, Y. Zhu, J. Gao, A. Garg, L. Fei-Fei, and S. Savarese, “Neural task programming: Learning to generalize across hierarchical tasks,” in (ICRA). IEEE, 2018, pp. 1–8.

[21] D.-A. Huang, S. Nair, D. Xu, Y. Zhu, A. Garg, L. Fei-Fei, S. Savarese, and J. C. Niebles, “Neural task graphs: Generalizing to unseen tasks from a single video demonstration,” in CVPR, 2019.

[22] L. Zhou, C. Xu, and J. J. Corso, “Towards automatic learning of procedures from web instructional videos,” arXiv preprint arXiv:1703.09788, 2017.

[23] A. Colas, S. Kim, F. Demoncourt, S. Gpeta, D. Z. Wang, and D. S. Kim, “Tutoriadvqa: Question answering dataset for tutorial videos,” LREC, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[27] M. Cuturi and M. Blondel, “Soft-dtw: a differentiable loss function for time-series,” International Conference on Machine Learning. PMLR, 2017, 2017.

[28] D. J. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series.” in KDD workshop. Seattle, WA, USA, 1994, pp. 359–370.

[29] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips,” in ICCV, 2019, pp. 2630–2640.

[30] L. v. d. Maaten and G. Hinton, “Visualizing data using T-sne,” Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[31] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in EMNLP, Oct. 2014.

[32] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[33] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv preprint arXiv:1412.3555, 2014.