Learning Causal Domain-Invariant Temporal Dynamics for Few-Shot Action Recognition

Yuke Li [*1] **Guangyi Chen** [*2] **Ben Abramowitz** [3] **Stefano Anzellotti** [1] **Donglai Wei** [1]

Abstract

Few-shot action recognition aims at quickly adapting a pre-trained model to the novel data with a distribution shift using only a limited number of samples. Key challenges include how to identify and leverage the transferable knowledge learned by the pre-trained model. We therefore propose **CDTD**, or **C**ausal **D**omain-Invariant **Te**mporal **D**ynamics for knowledge transfer. To identify the temporally invariant and variant representations, we employ the causal representation learning methods for unsupervised pertaining, and then tune the classifier with supervisions in next stage. Specifically, we assume the domain information can be well estimated and the pre-trained image decoder and transition models can be well transferred. During adaptation, we fix the transferable temporal dynamics and update the image encoder and domain estimator. The efficacy of our approach is revealed by the superior accuracy of **CDTD** over leading alternatives across standard few-shot action recognition datasets.

1. Introduction

Action recognition continues to be a potent and productive area of research. For instance, (Xing et al., 2023; Ahn et al., 2023; Zhou et al., 2023; Zhang et al., 2023) show great results in learning action representations with a large amount of labels. However, these supervised learning approaches may fall short when faced with the challenge of few-shot learning, where only a limited number of samples are available for novel action classes that exhibit significant distributional differences from the pre-training data (Cao et al., 2020; Perrett et al., 2021a; Thatipelli et al., 2022; Tseng et al., 2020; Luo et al., 2023).

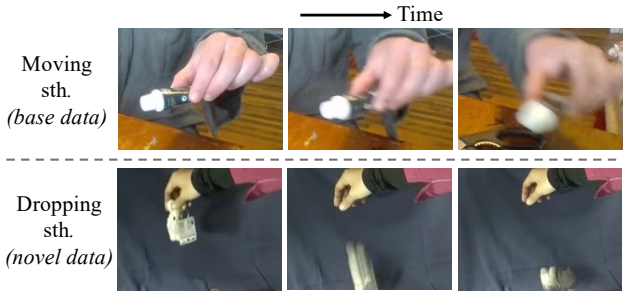


Figure 1: In the few-shot learning setting, what aspects can be effectively transferred from the base data to the novel data? Despite the different *temporal dynamics* in these two videos, the underlying physical laws are *domain-invariant*.

The efficient and effective few-shot action recognition requires solving two main problems. One entails identifying the transferable, temporal invariant knowledge that can be readily applied to new data. The other involves optimizing the non-transferable, temporal variant knowledge to facilitate swift updates, ensuring effective adaptation to new contexts. The existing literature relies on a large amount of labels and treats these two problems in various ways, based on different assumptions about transferability. For instance, one set of methods assumes all parameters should be fine-tuned on the new data, such as ActionCLIP (Wang et al., 2021b), ORViT (Herzig et al., 2022), and SViT (Ben Avraham et al., 2022). Other methods fix all parameters in the pre-trained model and learn new modules for data-efficient updates, such as the prompt learning methods VideoPrompt (Ju et al., 2022) and VL Prompting (Rasheed et al., 2023). By identifying and fixing domain-invariant parameters and only updating a small number of domain-specific parameters without introducing new modules during adaptation, **CDTD** offers a great improvement in adaptation efficiency over both of these approaches. We contrast our model also with prototypical network-based methods that focus only on the transferable part, learning the transferable knowledge via meta-learning and directly transferring a metric model for the recognition of new actions (Wang et al., 2021a; Zhu et al., 2021; Fu et al., 2020; Perrett et al., 2021b)

[*]Equal contribution [1]Boston College, Boston MA, USA [2]Carnegie Mellon University, Pittsburgh PA, USA; Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE [3]Tulane University, New Orleans LA, USA. Correspondence to: Donglai Wei <weidf@bc.edu>, Yuke Li <sunfreshing@whu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

In this paper, we propose a method, **CDTD**, that aims to learn domain-invariant features and optimizes the latter efficiently when faced with novel data. The recent minimal change principle (Kong et al., 2022) states when adapting a model to new data, one should make the smallest necessary changes to accommodate the new data. Taking inspiration from this principle, our key assumption is that domain information can be well estimated and temporal dynamic models are transferable. This assumption is inspired by the fact that the laws of physics remain constant across time, and therefore across all frames in any real-world video, and yet these laws are not directly observable. Any knowledge of the laws of physics learned by the pre-trained model should be transferable. We hypothesize that this is a general phenomenon, where temporal invariance in the relations between latent variables yields greater transferability.

Figure 1 exemplifies an example of our assumption by comparing "Moving Sth." to "Dropping Sth.". For both actions, the movement of a hand leads to the movement of an object. In each frame, we observe the position and orientation of the hand and the object, while their velocity and angle of motion are latent. Newton's laws of motion determine the acceleration (change in velocity) due to gravity in free fall, how the motion of the hand influences the motion of the object, etc. These relations are temporally invariant and can be transferred. Simultaneously, the two samples differ in aspects like the angle of motion of the hand. These aspects can change across the frames from a single sample and hence we expect them to be updated.

Specifically, we employ the temporal causal representation learning (Yao et al., 2022a; Chen et al., 2024) as the first unsupervised stage. Conversely, we assume that domain information is estimable and therefore does not require domain labels, making it a fully unsupervised learning process. Furthermore, we design the representation learning model with four modules: temporal dynamics generation, temporal dynamics transitions, visual encoder, domain estimator, and classifier. Classifier is learned after representation learning stage using labeled action classes while fixing the representation models. In the representation learning stage, we use the same temporal dynamic models for different domain and visual embeddings, by assuming that temporal dynamics are transferable. During adaption with few-shot examples, we fix the invariant temporal dynamics generation and transitions modules, and update the image encoder, domain estimator, and classifier.

Contributions. 1) We find that the temporal dynamics are transferable, and thus propose **CDTD**, a new framework with causal representation learning as a preliminary stage to distinguish invariant temporal dynamics and other variant knowledge. 2) We demonstrate the superior accuracy of **CDTD** over existing models across five standard benchmark datasets and validate our modeling assumptions through comprehensive ablation experiments. Our positive results are consistent with our central assumption that domain-invariance of the learned temporal dynamics suggests transferability to novel contexts.

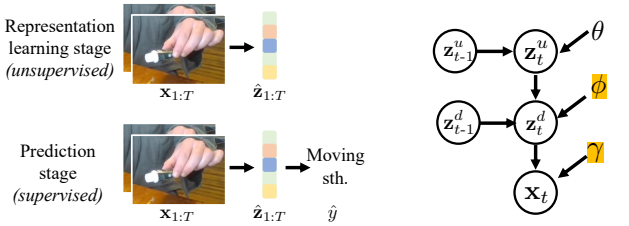
2. Related Work

Few-shot action recognition. Existing work on few-shot learning uses various approaches to deal with the temporal nature of action recognition (Finn et al., 2017). For instance, STRM (Thatipelli et al., 2022) introduces a spatio-temporal enrichment module to look at visual and temporal context at the patch and frame level. HyRSM (Wang et al., 2022) uses a hybrid relation model to learn relations within and across videos in a given few-shot episode. However, recent work has shown that these methods do not generalize well when the base and novel data have significant distribution disparities (Wang et al., 2023c; Samarasinghe et al., 2023).

The recent success of cross-modal vision-language learning has inspired works like ActionCLIP (Wang et al., 2021b), XCLIP is presented by the authors of (Ni et al., 2022), VideoPrompt (Ju et al., 2022), as well ViFi-CLIP is introduced in (Rasheed et al., 2023). Most of these methods apply transfer learning by adopting the pre-trained CLIP and adapting it for few-shot action recognition tasks with popular tuning strategies. For example, VL prompting fixes the backbone of CLIP and tunes the additional visual prompt (Jia et al., 2022) to adapt to the novel data. Nevertheless, if the CLIP model is pre-trained on a base dataset that is too dissimilar from the novel data, the issues from distribution disparity persist. Unlike previous work, our approach to few-shot action recognition leverages intuition from recent advancements in causal representation learning, providing an advantage for handling distribution disparities.

Domain-invariant feature learning. Domain generalization (DG), or out-of-distribution generalization, aims to enable domain adaptation when the base and novel data are not *i.i.d* and none of the novel data is available when training on the base data. Approaches for achieving DG include meta-learning (Li et al., 2018), data augmentation (Shorten & Khoshgoftaar, 2019), and domain-invariant feature learning (Matsuura & Harada, 2020; Lu et al., 2022). Domain-invariant feature learning naturally seeks to learn transferable knowledge that generalizes across domains but to improve the efficacy of few-shot learning we must also work with non-transferable knowledge.

The way **CDTD** distinguishes the transferable knowledge from the non-transferable knowledge is based on recent advances in causal representation learning on temporal data (Yao et al., 2022b; Feng et al., 2022; Yao et al., 2022a). Those works assume the disentanglement of the known



(a) Represent-predict pipeline (b) Temporal dynamics model

Figure 2: Represent-predict pipeline. (a) We first learn feature representation unsupervisedly for video frames and then supervisedly train the action prediction model. (b) In the "Represent" stage, we design a temporal dynamics model where parameters (ϕ , γ) are domain-invariant.

auxiliary domain index from the data generation process. In contrast, without known values for the auxiliary variable, we propose to estimate it instead.

3. Methodology

3.1. Few-shot Learning Framework Overview

Two-phase model training. A typical few-shot setting has two phases of model training on two sets of data: base and novel. The phase 1 training is on the base data $\mathcal{D}$, where the video action labels are from the set $\mathcal{C}_{base}$. The novel data consists of two parts, the support set $\mathcal{S}$ for updating the model and the query set $\mathcal{Q}$ for inference. Notably, there only exist limited samples for $\mathcal{S}$, and the video action labels are from the set $\mathcal{C}_{novel}$ that is disjoint from $\mathcal{C}_{base}$. The phase 2 training updates the model on the support set $\mathcal{S}$, aiming to improve the inference accuracy for novel classes in $\mathcal{Q}$.

Represent-predict pipeline. For each video clip with T frames, let $\mathbf{x}_{1:T}$ be the frame images and y be the action label. During each phase of model training, we follow the popular approach (Tong et al., 2022) to divide the pipeline into two stages (Fig. 2 (a)). During the representation learning stage, we train the model to extract feature $\hat{\mathbf{z}}_{1:T}$ from frame images $\mathbf{x}_{1:T}$ in an unsupervised manner. During the prediction stage, we fix the feature extractor and train the classifier to predict the action label $\hat{y}$. In this paper, we apply this framework and learn the causal representation in the first stage to better distinguish the invariant causal dynamics and other variant knowledge.

3.2. Causal Temporal Dynamics Model

Inspired by previous works on temporal causal modeling, we design a generative model based on the variational autoencoder (VAE) in the representation learning stage.

Data generation process. Each observation $\mathbf{x}_t$ is generated from a nonlinear mixing function $\mathbf{g}$ that maps latent variables $\mathbf{z}^d_t$ to $\mathbf{x}_t$, where $\mathbf{z}^d_t$ refers to the variables involved in the temporal dynamics. For every dimension $i \in 1, \ldots, m$ of the latent variable $\mathbf{z}^d_t$, $\mathbf{z}^d_t(i)$ is derived from a non-stationary, non-parametric time-delayed causal relation:

$$\underbrace{\mathbf{x}_t = \mathbf{g}(\mathbf{z}^d_t)}_{\text{Generation process}} ,$$
$$\underbrace{z^d_t(i) = f_i\left(\{z^d_{t'}(j)|z^d_{t'}(j) \in \mathbf{Pa}(z^d_t(i))\}, z^u_t(i), \epsilon_t(i)\right)}_{\text{Non-stationary non-parametric transition}} . \quad (1)$$

The non-stationary information is captured by the representation z^u_{it} and we assume that this information can be estimated from observed data. Here, we use the superscripts d and u to denote dynamic and domain information, respectively. The validity of this assumption is strong but reasonable since humans can typically estimate the domain information from the observed videos. We also verify the validity of this assumption in the experiments.

Probabilistic formulation. Without specifying, our notations follow the TDRL framework (Yao et al., 2022a). As shown in part (b) of Figure 2, we formulate the temporal dynamics model with $\{\mathbf{x}, \mathbf{z}^d, \mathbf{z}^u\}$. Specifically, for the convenience, we formulate the joint distribution with setting time-lag as 1:

$$p(\mathbf{x}_{1:T}, \mathbf{z}^d_{1:T}, \mathbf{z}^u_{1:T}) = p_\gamma(\mathbf{x}_1|\mathbf{z}^d_1)p_\phi(\mathbf{z}^d_1|\mathbf{z}^u_1)p(\mathbf{z}^u_1) \prod_{t=2}^{T} p_\gamma(\mathbf{x}_t|\mathbf{z}^d_t) \underbrace{p_\phi(\mathbf{z}^d_t|\mathbf{z}^d_{t-1}, \mathbf{z}^u_t)}_{\text{temporal dynamics}} p(\mathbf{z}^u_t|\mathbf{z}^u_{t-1}). \quad (2)$$

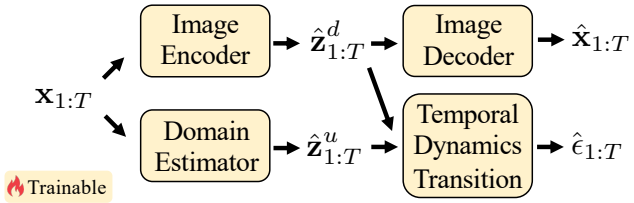
In the VAE framework, one can consider γ as the parameters for the image decoder (generation process) $p_\gamma(\mathbf{x}_t|\mathbf{z}^d_t)$, and ϕ as the parameters for the temporal dynamics transition $p_\phi(\mathbf{z}^d_t|\mathbf{z}^d_{t-1}, \mathbf{z}^u_t)$. To learn visual representation $\mathbf{z}^d_t$ and domain information $\mathbf{z}^u_t$, we also introduce two modules including image encoder $q_\omega(\mathbf{z}^d_t|\mathbf{x}_t)$ with parameters ω and domain estimator $q_\theta(\mathbf{z}^u_t|\mathbf{z}^u_{t-1}, \mathbf{x}_t)$ with parameters θ.

3.3. Network Designs

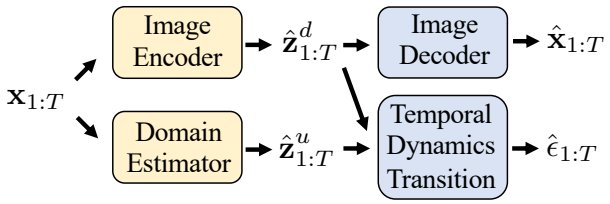
Image encoder $q_\omega(\mathbf{z}^d_t|\mathbf{x}_t)$. We denote the image encoder using $q_\omega(\hat{\mathbf{z}}^d_t|\mathbf{x}_t)$. We assume $\mathbf{z}^d_t$ is conditionally independent of all $\mathbf{z}^d_{t'}$ for $t' \neq t$ conditioned on $\mathbf{x}$, and therefore we can decompose the joint probability distribution of the posterior by $q_\omega(\hat{\mathbf{z}}^d|\mathbf{x}) = \prod_{t=1}^{T} q_\omega(\hat{\mathbf{z}}^d_t|\mathbf{x}_t)$. We choose to approximate q by an isotropic Gaussian characterized by mean μ_t and covariance σ_t. To learn the posterior we use an encoder composed of an MLP followed by leaky ReLU activation:

$$\hat{\mathbf{z}}^d_t \sim \mathcal{N}(\mu_t, \sigma_t), \quad \mu_t, \sigma_t = \text{LeakyReLU}(\text{MLP}(\mathbf{x}_t)). \quad (3)$$

Domain estimator $q_\theta(\mathbf{z}^u_t|\mathbf{z}^u_{t-1}, \mathbf{x}_t)$. Since we assume that $\mathbf{z}^u_t$ captures the domain variance for the transition of $\mathbf{z}_t$, $\mathbf{u}_t$



(a) Phase 1: Training on the base data



(b) Phase 2: Adaptation on the novel data

Figure 3: Our **CDTD** model for the representation learning stage. (a) During phase 1 on the base data, we train the whole causal temporal dynamics model. (b) During phase 2 on the novel data, we fine-tune both the image encoder and the domain estimator while freezing the image encoder and the temporal dynamicsOur **CDTD** model for the representation learning stage. (a) During phase 1 on the base data, we train the whole causal temporal dynamics model. (b) During phase 2 on the novel data, we fine-tune both the image encoder and the domain estimator while freezing the image encoder and the temporal dynamics transition module trained during phase 1. transition module trained during phase 1.

needs to be updated when adapting. In our task, we assume that the prior $p(\mathbf{z}^u_t)$ is not observed but we can estimate it. Specifically, we build the domain estimator on latent domain learning (Deecke et al., 2022). For each domain, we use a set of learnable gating functions, denoted as $A_1, A_2, \ldots, A_S$, to obtain $\hat{\mathbf{z}}^u_t$:

$$\hat{\mathbf{z}}^u_t = \text{GRU}(\hat{\mathbf{z}}^u_{t-1}, \sum_{\mathbf{s}=1}^{S} A_{\mathbf{s}}(\mathbf{x}_t)\text{Conv}_{\mathbf{s}}(\mathbf{x}_t)). \tag{4}$$

Here, Conv denotes a 1×1 convolution, enabling linear transformations for $\mathbf{x}$. The GRU models any temporal dependencies. We refer the readers to (Deecke et al., 2022) for more details.

Temporal Dynamics Transition $p_\phi(\mathbf{z}^d_t|\mathbf{z}^d_{t-1}, \mathbf{z}^u_t)$**.** Since the prior distribution of $p(\mathbf{z}^d_t|\mathbf{z}^d_{t-1}, \mathbf{z}^u_t)$ is not tractable, we employ a flow network to convert this computation with the distribution of the Gaussian noise ϵ_t. For $i-$th dimension of the noise vector, we formulate the prior module as $\hat{\epsilon}_t(i) = \hat{f}_i^{-1}(\hat{z}^d_t(i)|\hat{\mathbf{z}}^d_{t-1}, \hat{z}^u_t(i))$. Then the prior distribution of the i-th dimension of the temporal dynamics, $\hat{\mathbf{z}}^d_t(i)$, can be computed as

$$p_\phi\big(\hat{f}_i{}^{-1}(\hat{\mathbf{z}}^d_t(i)|\hat{\mathbf{z}}^d_{t-1}, \hat{\mathbf{z}}^u_t(i))\big)|\frac{\partial \hat{f}_i{}^{-1}}{\partial \hat{\mathbf{z}}^d_t(i)}|. \tag{5}$$

This flow model is built with the MLP layers. For a detailed derivation please refer to Appendix A.

Image decoder $p_\gamma(\mathbf{x}_t|\mathbf{z}^d_t)$**.** The image decoder pairs with our image encoder to generate an estimate of the action representations $\hat{\mathbf{x}}_t$ from the estimated latent variables $\hat{\mathbf{z}}^d_t$. We estimate $p_\gamma(\mathbf{x}_t|\mathbf{z}^d_t)$ using the decoder, which consists of a stacked MLP followed by leaky ReLU activation:

$$\hat{\mathbf{x}}_t = \text{LeakyReLU}(\text{MLP}(\hat{\mathbf{z}}^d_t)). \tag{6}$$

Classifier $p_\psi(y|\mathbf{z}^d_{1:T})$**.** We have two choices of ψ.

The first one is to the end of obtaining the standard one-hot embedding. We first concatenate $\hat{\mathbf{z}}^d_{1:T}$ along the time dimension and apply an MLP model for classification:

$$\hat{y} = \text{MLP}(\text{Concat}(\hat{\mathbf{z}}^d_{1:T})). \tag{7}$$

The alternative is the text embedding layer:

$$\hat{e} = \text{embed}(y) \tag{8}$$

We implement the text embedding layer using a transformer-based text encoder with eight attention heads, following the architecture used in (Rasheed et al., 2023).

3.4. Training Objectives

Representation learning stage. **CDTD** first employs the ELBO loss to learn γ, ϕ, θ,and ω aforementioned. The ELBO loss combines the reconstruction loss from using our decoder to estimate $\mathbf{x}_t$ with the KL-divergence between the posterior and prior:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\frac{1}{T}\sum_{t=1}^{T}\text{BCE}(\mathbf{x}_t, \hat{\mathbf{x}}_t)}_{\mathcal{L}_{\text{Recon}}} - \underbrace{\frac{1}{T}\sum_{t=1}^{T}\beta\mathbb{E}_{\hat{\mathbf{z}}^d_t\sim q_\omega}\left(\log q\left(\hat{\mathbf{z}}^d_t|\mathbf{x}_t\right) - \log p(\hat{\mathbf{z}}^d_t|\hat{\mathbf{z}}^d_{t-1}, \mathbf{z}^u_t)\right)}_{\mathcal{L}_{\text{KLD}}}. \tag{9}$$

Here, $\mathcal{L}_{\text{Recon}}$ measures the discrepancy between $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t$ using binary cross-entropy (BCE), and β is the hyperparameter to balance the two losses.

Prediction stage. Depending on the choice of the classifier, we use the corresponding loss function. The first choice is to use the CE loss using a one-hot encoding of y,

$$\mathcal{L}^{CE}_{\text{cls}} = -\mathbb{E}_{\hat{y}}\big(\text{one-hot}(y)\cdot\log(\text{softmax}(\hat{y}))\big). \tag{10}$$

Another choice is the NCE loss (Gutmann & Hyvärinen, 2010). For the k^{th} video sequence, let m be the indices of the action labels. We use the temporal pooling over $\hat{\mathbf{z}}^d$ obtain $\hat{\mathbf{z}}^d_{1:T,k}$.

$$\mathcal{L}^{NCE}_{\text{cls}} = -\sum_k \log \frac{\exp(\text{sim}(\hat{\mathbf{z}}^d_{1:T,k}, \hat{e}_k)/\tau)}{\sum_m \exp(\text{sim}(\hat{\mathbf{z}}^d_{1:T,k}, \hat{e}_m)/\tau)}, \quad (11)$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity between the text embedding of the action label and our learned $\hat{\mathbf{z}}^d_{1:T,k}$ and τ is the temperature parameter. In practice, we use Eq. 10 or Eq. 11 (Rasheed et al., 2023) for a fair comparison with previous methods, denoted as **CDTD**$_{CE}$ and **CDTD**$_{NCE}$ respectively.

3.5. Training, Adapting and Inference

The overall framework is illustrated in Figure 3. The model training involves two phases: in Phase 1, we train all modules with base data, while in Phase 2, we only adapt a part of modules with novel few-shot data.

Phase 1: Training on the base data $\mathcal{D}$. For the basic training, we first learn the causal temporal dynamics in an unsupervised way. Specifically, we jointly optimize the image encoder (ω), domain estimator (θ), dynamics transition (ϕ), and the image decoder (γ) with the ELBO loss in equation 9. Then, in the next stage, we fix the whole model and learn the classifier ψ with either CE or NCE losses in Eq. 10 and 11.

Phase 2: Adaptation on the novel data support set $\mathcal{S}$. Given the assumption that temporal dynamics are invariant on both base and novel data, we keep the learned temporal dynamics model unchanged in this phase. Specifically, in the representation stage for adaptation of the novel data, we only update the image encoder (ω) and domain encoder (θ). Then, we update the prediction model parameter (ψ) with a few shot labels.

Inference on the novel data query set $\mathcal{Q}$. To perform inference, we sample $\hat{\mathbf{z}}^d$ according to Eq. 3 using our image encoder, and we either choose the maximum value (highest probability) from the prediction $\hat{y}$ or choose the label whose text embedding maximizes cosine similarity.

4. Experiments

4.1. Experimental Setup

We carry out two types of few-shot learning experiments; all-way-k-shot and 5-way-k-shot. (Ju et al., 2022) proposes the setting of all-way-k-shot: we try to classify all action classes in the class for the novel dataset ($\mathcal{C}_{novel}$), while in 5-way-k-shot learning we only try to estimate 5 label classes at a time in a series of trials. The number of shots (k) refers to the number of training samples in $\mathcal{S}$ available for each action label. Once we partition our data into base set $\mathcal{D}$ and novel set $\mathcal{S} \cup \mathcal{Q}$, the number k determines how many samples for each action class we choose for $\mathcal{S}$ and the rest are used for the query set $\mathcal{Q}$.

Datasets. In this work, we conduct experiments on five datasets: 1. Something-Something v2 (SSv2) is a dataset containing 174 action categories of common human-object interactions; 2. Something-Else (Sth-Else) exploits the compositional structure of SSv2, where a combination of a verb and a noun defines an action; 3. HMDB-51 contains 7k videos of 51 categories; 4. UCF-101 covers 13k videos spanning 101 categories; 5. Kinetics covers around 230k 10-second video clips sourced from YouTube. For the experiments that perform training and testing on the Sth-Else dataset, we use the official split of data (Materzynska et al., 2020; Herzig et al., 2022; Ben Avraham et al., 2022). Similarly, for experiments using other datasets, we split the data into $\mathcal{D}$, $\mathcal{S}$, and $\mathcal{Q}$ following the prior work we use as benchmarks, as described late. The other datasets we use for novel data are SSv2 (Goyal et al., 2017), SSv2-small (Zhu & Yang, 2018), HMDB-51 (Kuehne et al., 2011), and UCF-101 (Soomro et al., 2012). **Evaluation metrics.** In all experiments, we compare the Top-1 accuracy, i.e., the maximum accuracy on any action class, of **CDTD** against leading benchmarks for few-shot action recognition. The results of the top-performing model are given in **bold**, and the second-best are underlined.

Implementation details. In each experiment, there is a backbone used to extract the action representations $\mathbf{x}_t$ used for training, and in all experiments, the backbone is either ResNet-50 (He et al., 2016) or ViT-B/16 (Radford et al., 2021) trained on the base data set $\mathcal{D}$, where the choice of backbone matches what was used in the benchmark experiments. We use the AdamW optimizer (Loshchilov & Hutter, 2019) and cosine annealing to train our network with a learning rate initialized at 0.002 and weight decay of 10^{-2}. For all video sequences we use $T = 16$ uniformly selected frames. To compute the ELBO loss, we choose $\beta = 0.02$ to balance the reconstruction loss and KL-divergence. Also, we set $\tau = 0.07$ for the NCE loss. Regarding the hyperparameters, we set $d = 12$ in Eq. 1, and $S = 35$ in Eq. 4. The hyperparameter analysis is reported in Table 5. For a detailed summary of our network architectures see Table 9 in the appendix. Our models are implemented using PyTorch, and experiments are conducted on four Nvidia GeForce 2080Ti graphics cards.

4.2. Benchmark Results

In the first part of our experiment, we compare the results from **CDTD** against state-of-the-art few-shot action recog-

Table 1: Benchmark results on the Sth-Else dataset with the same ViT-B/16 backbone but two different loss functions.

	Loss	Sth-Else	
		5-shot	10-shot
ORViT (Herzig et al., 2022)	CE	33.3	40.2
SViT (Ben Avraham et al., 2022)		34.4	42.6
CDTD$_{CE}$ (ours)		**37.6**	**44.0**
ViFi-CLIP (Rasheed et al., 2023)	NCE	44.5	54.0
VL Prompting (Rasheed et al., 2023)		44.9	58.2
CDTD$_{NCE}$ (ours)		**48.5**	**63.9**

nition methods. To obtain fair comparisons, we compare **CDTD** with the leading approaches using identical backbones and classification loss on each dataset.

Sth-Else Experiments. Table 1 shows two experiments against benchmarks for all-way-k-shot learning using the ViT-B/16 backbone for the Sth-Else dataset. We compare **CDTD**$_{CE}$ to leading benchmarks on Sth-Else that use cross-entropy loss, ORViT and SViT (Herzig et al., 2022; Ben Avraham et al., 2022). We also compare **CDTD**$_{NCE}$ to the state-of-the-art methods that employ contrastive learning, ViFi-CLIP, and VL-Prompting (Rasheed et al., 2023).

It is evident that **CDTD**$_{NCE}$ and **CDTD**$_{CE}$ had the highest accuracy in both experiments. **CDTD** outperforms all four leading benchmarks by amounts ranging from 1.4 to 5.7 percentage points. For this dataset we see the greatest improvement for $k = 5$, improving from 34.4 to 37.6 and from 44.9 to 48.5 over the leading benchmarks, and with improvements from 58.2 to 63.9 for $k = 10$.

Figure 5 further shows the number of parameters updated during transfer. Our **CDTD** necessitates tuning the fewest parameters among all the approaches when adapting from the base data to the novel data, bringing great parameter efficiency. Notably, ViFi-CLIP, ORViT, and SViT require an order of magnitude more parameters to be tuned for adaptation to novel data compared to **CDTD**. Similarly, the VPT (Jia et al., 2022) of CLIP needed for VL Prompting results in more than **40 million** parameters needing updating.

Fig. 4 shows that **CDTD** outperforms VL-Prompting on the majority of action classes. Notably, **CDTD** exceeds the VL prompt in 72 out of 86 action classes, such as "plugging something into something" and "spinning so it continues spinning." We also note significant improvements in other categories, including "approaching something with something" and "pulling two ends of something but nothing happens". However, **CDTD** does have some limitations. Specifically, VL-Prompting surpasses our method in label classes like "burying something in something" and "lifting a surface with something on it".

All-way-k-shot Experiments. Table 2 show the experimental results using the ViT-B/16 backbone comparing **CDTD**$_{NCE}$ to leading benchmarks for all-way-k-shot learning for $k \in \{2, 4, 8, 16\}$. For all-way-k-shot learning, our benchmarks are: ActionCLIP (Wang et al., 2021b), XCLIP (Ni et al., 2022), VideoPrompt (Ju et al., 2022), VL Prompting and ViFi-CLIP (Rasheed et al., 2023), VicTR (Kahatapitiya et al., 2023) and VideoMAE (Tong et al., 2022).

We follow (Wang et al., 2021b; Ni et al., 2022; Ju et al., 2022; Rasheed et al., 2023) in using Kinetics-400 (K-400) from (Carreira & Zisserman, 2017) as the base dataset $\mathcal{D}$ and repeat the experiment using novel data from the SSV2, HMDB-51, and UCF-101 datasets. We can observe that **CDTD**$_{NCE}$ had the highest accuracy in 11 out of 12 of these experiments and had the second highest accuracy in the remaining experiment trailing by only 0.4.

5-way-k-shot Experiments. Table 3 shows the results of experiments comparing **CDTD**$_{CE}$ to leading benchmarks for 5-way-k-shot learning. All models are trained on K-400 as the base data using the ResNet-50 backbone. In each trial, we select 5 action classes at random and update our model using only k samples for each of those classes to form $\mathcal{S}$, and the remaining novel data compose $\mathcal{Q}$. To ensure the statistical significance, we conduct 1000 trials with random samplings for selecting the action classes in each trial by following (Wang et al., 2023c), and report the mean accuracy as the final result. Our leading benchmarks are: OTAM (Cao et al., 2020), TRX (Perrett et al., 2021a), STRM (Thatipelli et al., 2022), DYDIS (Islam et al., 2021), STARTUP (Phoo & Hariharan, 2021), and SEEN (Wang et al., 2023c). **CDTD**$_{CE}$ has the highest accuracy in 7 out of 8 experiments, by a margin ranging from 0.6 to 5.3, and has the second highest accuracy on the remaining experiment, trailing by only 0.1.

4.3. Ablation Studies

Components of CDTD. In this section, we demonstrate the contribution of each component of our **CDTD** model. We proceeded by comparing our **CDTD** model to four simpler baselines: w/o temporal dynamic transition, w/o domain encoder, w/o temporal modeling, and the full model with all components fine-tuned **CDTD**-FT. The model w/o temporal dynamic transition removes the temporal dynamic transition and domain encoder. The posterior is regularized by KL-divergence with the standard normal distribution. The model w/o temporal modeling assumes $\mathbf{z}_{t-l} = \mathbf{z}_t$ for all t. Per time step, the posterior in w/o temporal baseline regularizes the prior by KL-divergence independently from other time steps. The model w/o domain encoder removes the domain encoder module so only the image encoder ω and classifier ϕ are updated during adaptation. For **CDTD**-FT, we also finetune the temporal dynamic transition and image decoder rather than fixing them during adaptation. **CDTD**-UG updates the temporal dynamic transition module during

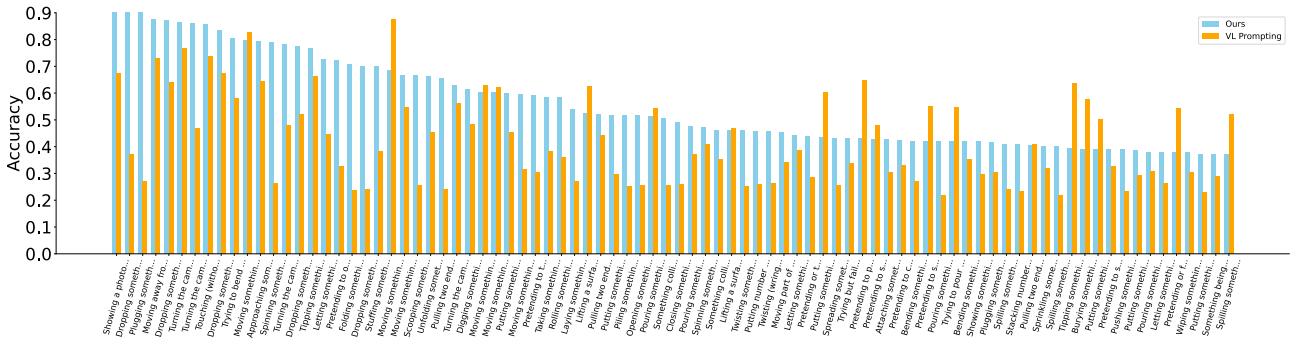


Figure 4: Comparing performance of **CDTD**$_{NCE}$ (blue) against VL-Prompting (orange) across all action classes on the Sth-Else dataset.

Table 2: Benchmark results for the all-way-k-shot learning setting. All models use ViT-B/16 backbone and the NCE loss.

	SSv2				HMDB-51				UCF-101			
	2-shot	4-shot	8-shot	16-shot	2-shot	4-shot	8-shot	16-shot	2-shot	4-shot	8-shot	16-shot
XCLIP(Ni et al., 2022)	3.9	4.5	6.8	10.0	53.0	57.3	62.8	64.0	70.6	71.5	73.0	91.4
ActionCLIP(Wang et al., 2021b)	4.1	5.8	8.4	11.1	47.5	57.9	57.3	59.1	70.6	71.5	73.0	91.4
VicTR(Kahatapitiya et al., 2023)	4.2	6.1	7.9	10.4	60.0	63.2	66.6	70.7	87.7	92.3	93.6	95.8
VideoPrompt(Ju et al., 2022)	4.4	5.1	6.1	9.7	39.7	50.7	56.0	62.4	71.4	79.9	85.7	89.9
ViFi-CLIP(Rasheed et al., 2023)	6.2	7.4	8.5	12.4	57.2	62.7	64.5	66.8	80.7	85.1	90.0	92.7
VL Prompting(Rasheed et al., 2023)	6.7	7.9	10.2	13.5	63.0	65.1	69.6	72.0	**91.0**	93.7	95.0	96.4
VideoMAE (Tong et al., 2022)	8.2	10.0	15.1	18.2	63.7	69.4	70.9	75.3	**91.0**	94.1	94.8	97.7
CDTD$_{NCE}$ **(ours)**	**9.5**	**11.6**	**14.8**	**19.5**	**65.8**	**70.2**	**72.5**	**77.9**	90.6	**94.7**	**96.2**	**98.5**

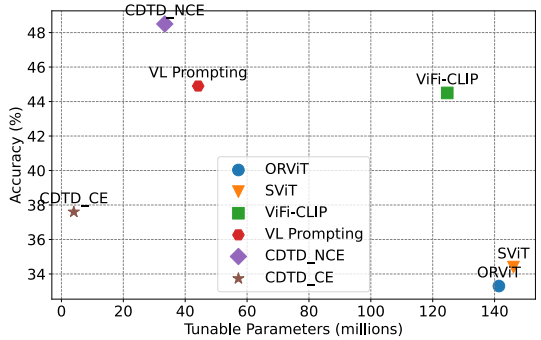


Figure 5: The comparison of the parameters that need to be updated during adapting on **Sth-Else dataset**.

adaptation, while the temporal dynamic generation module remains fixed. **CDTD**-UT updates all modules during the adaptation process.

We summarize the results of our ablation study in Table 4. First, notice that the accuracy of 45.1 by the w/o temporal modeling is already better than the 44.5 and 44.9 achieved by ViFi-CLIP and VL-prompting, respectively (Rasheed et al., 2023) (Table 1). This demonstrates the superior transportability of our method. Moreover, **CDTD** performs even better because it captures the temporal relations of the latent causal variables, and $\boldsymbol{\theta}$ and ω help capture the distribution shift across data sets. The superiority of **CDTD** over both w/o causal and w/o temporal demonstrates the advantage of modeling the action generation process.

Hyperparamter sensitivity. **CDTD** has three hyperparameters: the numbers of latent variables d in Eq. 1, and the number of latent domains S in Eq. 4. We also extend the time lags by choosing $l = 1$ to $l = 3$ in Eq. 2 We report the results varying these hyperparameters in Table 5 using **CDTD**$_{NCE}$ on the Sth-Else dataset. We vary one variable at a time, keeping the others constant.

Based on our ablation experiments in Table 5 we fix $d = 12$, $l = 1$, and $S = 35$ in all of our experiments unless otherwise stated. **CDTD**$_{NCE}$ performs better with bigger d, verifying that the latent causal variables facilitate action representation. We choose $d = 12$ because beyond this point the performance gain becomes marginal and the computational cost spikes. If we allow the parents of each $\mathbf{z}$ to be further back in time ($l > 1$), allowing causal relationships between non-consecutive time steps, the change in accuracy is negligible. Therefore, we choose that the parents of each $\mathbf{z}$ are only in the previous time step ($l = 1$). Finally, a consistent improvement is observed when increasing S from 10 to 50. However, considering the increased computational cost associated with the linear transformations and 1×1 convolutions involved in Eq. 4, we have opted for $S = 35$ as the standard across our experiments.

Table 3: Benchmark results for the 5-way-k-shot learning setting. All models use the ResNet-50 backbone and the CE loss.

	1-shot				5-shot			
	UCF-101	HMDB-51	SSv2	SSv2-small	UCF-101	HMDB-51	SSv2	SSv2-small
OTAM(Cao et al., 2020)	50.2	34.4	24.0	22.4	61.7	41.5	27.1	25.8
TRX(Perrett et al., 2021a)	47.1	32.0	23.2	22.9	66.7	43.9	27.9	26.0
STRM(Thatipelli et al., 2022)	49.2	33.0	23.6	22.8	67.0	45.2	28.7	26.4
DYDIS(Islam et al., 2021)	63.4	35.2	25.3	24.8	77.5	50.8	29.3	27.2
STARTUP(Phoo & Hariharan, 2021)	65.4	35.5	25.1	25.0	79.5	50.4	31.3	28.7
SEEN(Wang et al., 2023c)	64.8	35.7	26.1	25.3	**79.8**	51.1	34.4	29.3
CDTD$_{CE}$ (ours)	**66.0**	**37.6**	**31.4**	**28.5**	79.7	**54.4**	**37.0**	**32.8**

Table 4: Ablation study results on the effect of each component on the Sth-Else dataset.

	5-shot	10-shot
w/o temporal dynamic transition	41.0	44.3
w/o domain encoder	42.8	50.6
w/o temporal modeling	45.1	54.4
CDTD$_{NCE}$-FT	44.8	58.3
CDTD$_{NCE}$-UG	46.0	61.3
CDTD$_{NCE}$-UT	46.2	60.8
CDTD$_{NCE}$	**48.5**	**63.9**

Table 5: Ablation study results on the hyperparameter sensitivity on the Sth-Else dataset.

	5-shot	10-shot
$d = 4$	44.2	49.0
$d = 8$	46.0	55.7
$d = 12$	48.5	63.9
$d = 16$	**48.9**	**65.1**
$l = 1$	48.5	63.9
$l = 2$	48.5	64.0
$l = 3$	**48.8**	**64.6**
$S = 10$	42.5	51.1
$S = 20$	46.4	59.8
$S = 35$	48.5	63.9
$S = 50$	**49.1**	**65.0**

4.4. Discussion

The distributional disparities between the base and novel data. To illustrate the distributional differences between the base and novel data sets, we employed a pre-trained ViT-B/16 backbone to extract action representations from each. As demonstrated in Figure 6a, the Sth-Else dataset is characterized by significant distributional disparities between the base and novel data. This aspect notably intensifies the complexity of the few-shot action recognition task, highlighting the challenges inherent in adapting models from the base data to the novel data.

Validating of our assumptions. To validate our theoretical assumptions, we independently trained two **CDTD** models on the base and novel data sets, respectively.

We visualized several examples of the temporal dynamic transition module through pairplots representing the relationship between the latent variables across three consecutive time steps. We assume that if the temporal dynamic transition is consistent across the data, there should be a significant alignment between these variables. This alignment is indeed observed in Figure 6d even under the severe data distributional disparities in Figure 6a. Additionally, the impact of $\mathbf{u}$ on the temporal dynamics is shown in Figure 6c. The better accuracy scores from Table 1 to 4 support our assumption that $\mathbf{u}$ captures the distributional discrepancies.

Figure 6b similarly displays the pairplots between $\mathbf{x}$ and $\mathbf{z}$ for all time steps, illustrating the relationship between observed data and latent variables, i.e., temporal dynamic generation. The apparent overlap suggests that our assumption that holding the temporal dynamic generation fixed holds true.

We also empirically validate our assumptions on SSv2, HMDB-51, and UCF-101 datasets in Figures 7 in the appendix. We would like to highlight the consistency of the evidence regardless of the size of distributional disparities between the base and novel data.

5. Conclusion

We introduced the Domain-Invariant Temporal Dynamics (**CDTD**) framework for few-shot action recognition. Following the common two-stage training strategy, our framework involves unsupervised feature learning followed by supervised label prediction. In the unsupervised feature learning stage, we incorporate a learnable domain index to help the learned temporal dynamics modules to be more invariant to domain shift. These modules remain fixed during few-shot adaptation. Empirically, we validate the effectiveness of this assumption by achieving new state-of-the-art results on various standard action recognition datasets with minimal parameter updates during the adaptation process.

Limitations. This paper demonstrates the effectiveness of the learned temporal dynamics empirically. However, we believe that the effectiveness of any adaptation strategy from the base to novel data should be theoretically grounded, particularly concerning the generalization bound. Currently,

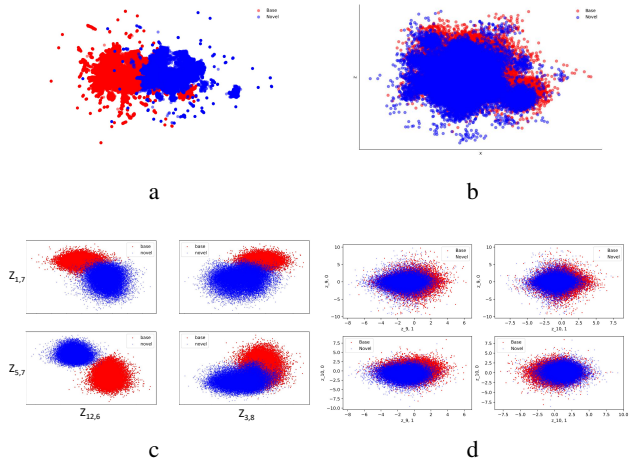


Figure 6: Visualizations to substantiate our hypotheses on the Sth-Else dataset. Red: base data. Blue: novel data. (a) This UMap visualization contrasts the action features extracted using a static ViT-B/16 backbone (Radford et al., 2021) from the base and novel data within the Sth-Else dataset. The stark distributional disparities underscore the challenges inherent in few-shot action recognition learning. (b) A pairplot of UMap projections for the action feature embeddings, $\mathbf{x}$, and the latent variables, $\mathbf{z}$, derived from two instances of our model's temporal dynamics generation module, trained separately on the base and novel data. (c) A pairplot comparing values of $\mathbf{z}_{n,t}$ from two models' temporal dynamics transition modules, each integrating the $\mathbf{u}$ from the base and novel data, respectively. (d) The temporal dynamics of latent variables using a consistent $\mathbf{z}_t^u$ across both base and novel data. The close alignment observed across datasets implies that the temporal dynamics transition modules are consistent, confirming the stability of these functions regardless of the data source.

this aspect is not adequately addressed in our work. Such an investigation could provide deeper insights into which components of a model should remain fixed versus which should be updated during adaptation. Extending **CDTD** to address instantaneous causal relations and providing a generalization bound are clear directions for future work. Lastly, we mention that exploring the merits of **CDTD** for LLM agents might be an interesting direction (Yao et al., 2023; Wang et al., 2023a).

Acknowledgements

This work is supported by National Science Foundation, Division of Information and Intelligent Systems (No. 2239688).

Impact Statement

The challenge of recognizing human actions from video sequences is currently at the frontier of machine learning. It has many real-world applications ranging from security and behavior analysis to gaming and human-robot interaction. One of the major difficulties identified in the literature on action recognition is the perennial dearth of labeled data for certain action classes due to the sheer number of possible actions. Few-shot learning is one of the primary approaches to this problem of scarcity. However, few-shot learning struggles when there is a significant distribution shift between the original data used for the pre-trained model and the novel data from which one wishes to recognize actions. Our approach develops a model of temporal relations that distinguishes the parts of the model that transfer to the novel context from those that do not. By limiting the number of parameters that must be fine-tuned to the novel data we have developed a model that can be adapted efficiently and provides higher accuracy than the previous leading benchmarks. Our hope is that future work will extend our approach with increasingly realistic temporal relation models that enable action recognition with higher accuracy. While we recognize that the field of action recognition broadly does pose risks, including related to automated surveillance, our work does not present any significant new risk to the field. Ultimately, this paper presents work whose goal is to advance the field of Machine Learning, and particularly action recognition. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ahn, D., Kim, S., Hong, H., and Ko, B. C. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3330–3339, 2023.

Ben Avraham, E., Herzig, R., Mangalam, K., Bar, A., Rohrbach, A., Karlinsky, L., Darrell, T., and Globerson, A. Bringing image scene structure to video via frame-clip consistency of object tokens. *Advances in Neural Information Processing Systems*, 35:26839–26855, 2022.

Cao, K., Ji, J., Cao, Z., Chang, C.-Y., and Niebles, J. C. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10618–10627, 2020.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Chen, G., Shen, Y., Chen, Z., Song, X., Sun, Y., Yao, W., Liu, X., and Zhang, K. Caring: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

Deecke, L., Hospedales, T., and Bilen, H. Visual representation learning over latent domains. In *International Conference on Learning Representations*, 2022.

Feng, F., Huang, B., Zhang, K., and Magliacane, S. Factored adaptation for non-stationary reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 31957–31971, 2022.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Fu, Y., Zhang, L., Wang, J., Fu, Y., and Jiang, Y.-G. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1142–1151, 2020.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Herzig, R., Ben-Avraham, E., Mangalam, K., Bar, A., Chechik, G., Rohrbach, A., Darrell, T., and Globerson, A. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3148–3159, June 2022.

Islam, A., Chen, C.-F. R., Panda, R., Karlinsky, L., Feris, R., and Radke, R. J. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *Advances in Neural Information Processing Systems*, 34: 3584–3595, 2021.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Ju, C., Han, T., Zheng, K., Zhang, Y., and Xie, W. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pp. 105–124. Springer, 2022.

Kahatapitiya, K., Arnab, A., Nagrani, A., and Ryoo, M. S. Victr: Video-conditioned text representations for activity recognition. *arXiv preprint arXiv:2304.02560*, 2023.

Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. Partial disentanglement for domain adaptation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11455–11472. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kong22a.html.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lu, W., Wang, J., Li, H., Chen, Y., and Xie, X. Domain-invariant feature exploration for domain generalization. *Transactions on Machine Learning Research*, 2022.

Luo, X., Wu, H., Zhang, J., Gao, L., Xu, J., and Song, J. A closer look at few-shot classification again. *arXiv preprint arXiv:2301.12246*, 2023.

Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., and Darrell, T. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1049–1059, 2020.

Matsuura, T. and Harada, T. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11749–11756, 2020.

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., and Ling, H. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2022.

Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., and Damen, D. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 475–484, 2021a.

Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., and Damen, D. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 475–484, 2021b.

Phoo, C. P. and Hariharan, B. Self-training for few-shot transfer across extreme task differences. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O3Y56aqpChA.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rasheed, H., Khattak, M. U., Maaz, M., Khan, S., and Khan, F. S. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6545–6554, 2023.

Samarasinghe, S., Rizve, M. N., Kardan, N., and Shah, M. Cdfsl-v: Cross-domain few-shot learning for videos. *arXiv preprint arXiv:2309.03989*, 2023.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Thatipelli, A., Narayan, S., Khan, S., Anwer, R. M., Khan, F. S., and Ghanem, B. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19958–19967, 2022.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Tseng, H.-Y., Lee, H.-Y., Huang, J.-B., and Yang, M.-H. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2020.

Wang, J., Wang, Y., Liu, S., and Li, A. Few-shot fine-grained action recognition via bidirectional attention and contrastive meta-learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 582–591, 2021a.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023a.

Wang, M., Xing, J., and Liu, Y. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021b.

Wang, X., Zhang, S., Qing, Z., Tang, M., Zuo, Z., Gao, C., Jin, R., and Sang, N. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19948–19957, 2022.

Wang, X., Zhang, S., Qing, Z., Gao, C., Zhang, Y., Zhao, D., and Sang, N. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18011–18021, 2023b.

Wang, X., Zhang, S., Qing, Z., Lv, Y., Gao, C., and Sang, N. Cross-domain few-shot action recognition with unlabeled videos. *Computer Vision and Image Understanding*, pp. 103737, 2023c.

Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., and Jiang, Y.-G. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18816–18826, 2023.

Yao, W., Chen, G., and Zhang, K. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022a.

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b.

Yao, W., Heinecke, S., Niebles, J. C., Liu, Z., Feng, Y., Xue, L., Murthy, R., Chen, Z., Zhang, J., Arpit, D., et al. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*, 2023.

Zhang, C., Gupta, A., and Zisserman, A. Helping hands: An object-aware ego-centric video recognition model. *arXiv preprint arXiv:2308.07918*, 2023.

Zheng, S., Chen, S., and Jin, Q. Few-shot action recognition with hierarchical matching and contrastive learning. In *European Conference on Computer Vision*, pp. 297–313. Springer, 2022.

Zhou, X., Arnab, A., Sun, C., and Schmid, C. How can objects help action recognition? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2353–2362, 2023.

Zhu, L. and Yang, Y. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 751–766, 2018.

Zhu, X., Toisoul, A., Perez-Rua, J.-M., Zhang, L., Martinez, B., and Xiang, T. Few-shot action recognition with prototype-centered attentive learning. *arXiv preprint arXiv:2101.08085*, 2021.

A. Transition Prior Likelihood Derivation

Consider a paradigmatic instance of latent causal processes. In this case, we are concerned with two time-delayed latent variables, namely, $\mathbf{z}_t = [\mathbf{z}_{1,t}, \mathbf{z}_{2,t}]$. Crucially, there is no inclusion of $\mathbf{u}$. We set time lag is defined as 1 for simplicity. This implies that each latent variable, $\mathbf{z}_{n,t}$, is formulated as $\mathbf{z}_{n,t} = f_n(\mathbf{z}_{t-1}, \epsilon_{n,t})$, where the noise terms, $\epsilon_{n,t}$, are mutually independent. To represent this latent process more succinctly, we introduce a transformation map, denoted as f. It's worth noting that in this context, we employ an overloaded notation; specifically, the symbol f serves dual purposes, representing both transition functions and the transformation map.

$$\begin{bmatrix} \mathbf{z}_{1,t-1} \\ \mathbf{z}_{2,t-1} \\ \mathbf{z}_{1,t} \\ \mathbf{z}_{2,t} \end{bmatrix} = \mathbf{f} \left(\begin{bmatrix} \mathbf{z}_{1,t-1} \\ \mathbf{z}_{2,t-1} \\ \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \right). \tag{12}$$

By leveraging the change of variables formula on the map $\mathbf{f}$, we can evaluate the joint distribution of the latent variables $p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}, \mathbf{z}_{1,t}, \mathbf{z}_{2,t})$ as:

$$p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}, \mathbf{z}_{1,t}, \mathbf{z}_{2,t}) = p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}, \epsilon_{1,t}, \epsilon_{2,t}) / \left| \det \mathbf{J}_{\mathbf{f}} \right|, \tag{13}$$

where $\mathbf{J}_{\mathbf{f}}$ is the Jacobian matrix of the map $\mathbf{f}$, which is naturally a low-triangular matrix:

$$\mathbf{J}_{\mathbf{f}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{\partial \mathbf{z}_{1,t}}{\partial \mathbf{z}_{1,t-1}} & \frac{\partial \mathbf{z}_{1,t}}{\partial \mathbf{z}_{2,t-1}} & \frac{\partial \mathbf{z}_{1,t}}{\partial \epsilon_{1,t}} & 0 \\ \frac{\partial \mathbf{z}_{2,t}}{\partial \mathbf{z}_{1,t-1}} & \frac{\partial \mathbf{z}_{2,t}}{\partial \mathbf{z}_{2,t-1}} & 0 & \frac{\partial \mathbf{z}_{2,t}}{\partial \epsilon_{2,t}} \end{bmatrix}.$$

Given that this Jacobian is triangular, we can efficiently compute its determinant as $\prod_n \frac{\partial \mathbf{z}_{n,t}}{\partial \epsilon_{n,t}}$. Furthermore, because the noise terms are mutually independent, and hence $\epsilon_{n,t} \perp \epsilon_{l,t}$ for $m \neq n$ and $\epsilon_t \perp \mathbf{z}_{t-1}$, we can write Eq. 13 as:

$$\begin{aligned} p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}, \mathbf{z}_{1,t}, \mathbf{z}_{2,t}) &= p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}) \times p(\epsilon_{1,t}, \epsilon_{2,t}) / \left| \det \mathbf{J}_{\mathbf{f}} \right| \quad \text{(because } \epsilon_t \perp \mathbf{z}_{t-1}\text{)} \\ &= p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}) \times \prod_n p(\epsilon_{n,t}) / \left| \det \mathbf{J}_{\mathbf{f}} \right| \quad \text{(because } \epsilon_{1,t} \perp \epsilon_{2,t}\text{)} \end{aligned} \tag{14}$$

By eliminating the marginals of the lagged latent variable $p(\mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1})$ on both sides, we derive the transition prior likelihood as:

$$p(\mathbf{z}_{1,t}, \mathbf{z}_{2,t} | \mathbf{z}_{1,t-1}, \mathbf{z}_{2,t-1}) = \prod_n p(\epsilon_{n,t}) / \left| \det \mathbf{J}_{\mathbf{f}} \right| = \prod_n p(\epsilon_{n,t}) \times \left| \det \mathbf{J}_{\mathbf{f}}^{-1} \right|. \tag{15}$$

Let $\{f_n^{-1}\}_{n=1,2,3...}$ be a set of learned inverse dynamics transition functions that take the estimated latent causal variables in the fixed dynamics subspace and lagged latent variables, and output the noise terms, i.e., $\hat{\epsilon}_{n,t} = f_n^{-1}\left(\hat{\mathbf{z}}_{n,t}, \mathrm{Pa}(\hat{\mathbf{z}}_{n,t})\right)$.

The differences of our model from Eq. 15 are that the learned inverse dynamics transition functions take $\mathbf{z}_t^u$ as input arguments to out the noise terms, i.e, $\hat{\epsilon}_{n,t} = f_n^{-1}\left(\hat{\mathbf{z}}_{n,t}, \mathrm{Pa}(\hat{\mathbf{z}}_{n,t}), \mathbf{u}_t\right)$.

$$\log p\left(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-l}, \mathbf{z}_t^u\right) = \sum_{n=1}^{N} \log p(\hat{\epsilon}_{n,t} | \mathbf{z}_t^u) + \sum_{n=1}^{N} \log \left| \frac{\partial f_n^{-1}}{\partial \hat{\mathbf{z}}_{n,t}} \right| \tag{16}$$

B. Assumption validation on SSv2, HMDB51 and UCF101 dataset

Figure 7 showcases the similar motivating examples of a few-shot action recognition task with **CDTD** in using SSv2, HMDB-51 and UCF-101 as novel data, while the K-400 serves as base data. Please see the caption for the detailed explanations.

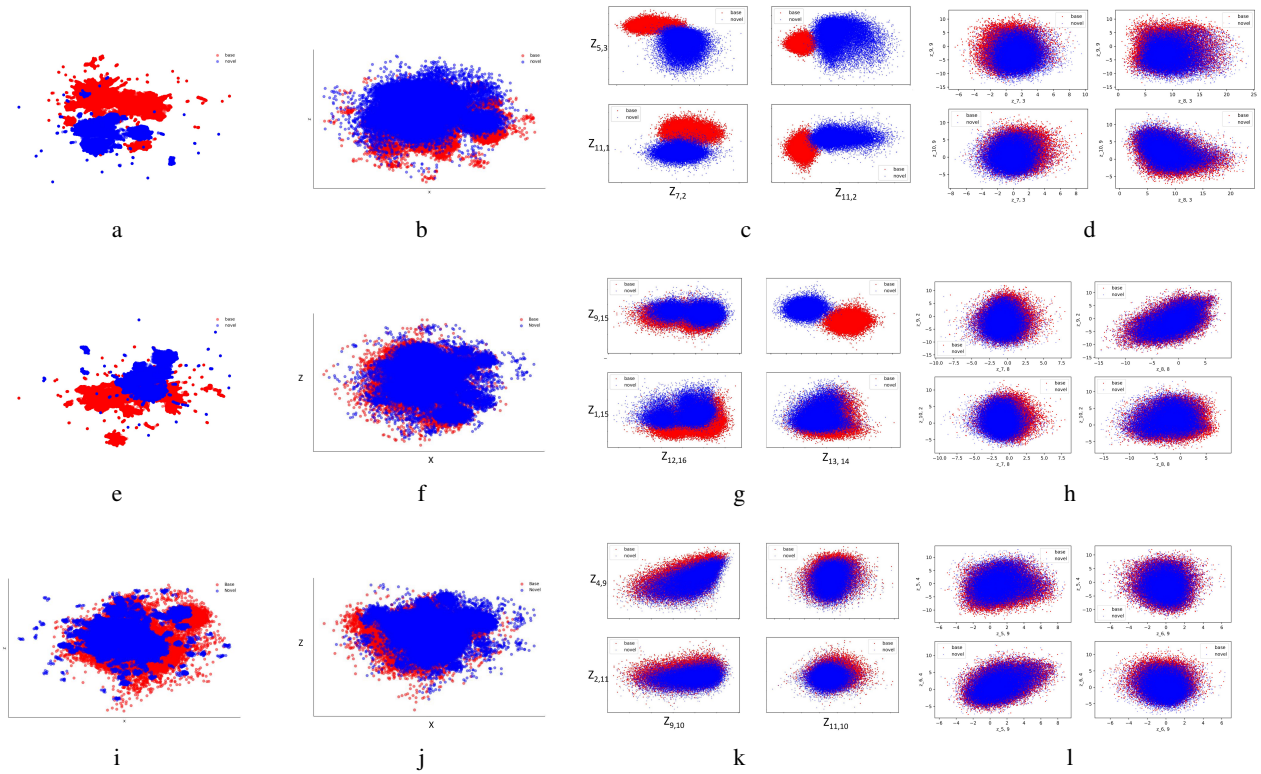


Figure 7: Visualizations validating hypotheses on SSv2, HMDB-51 and UCF-101 datasets (from top row to bottom row), respectively. Red: base data (K-400), blue: novel data. (a,e,i) UMap of action embeddings. (b,f,j) Pairplots of UMap projections for embeddings and latent variables. (c,g,k) Pairplots of latent variables from temporal invariance functions. (d,h,l) Temporal dynamics of latent variables with consistent $\mathbf{z}_t^u$.

C. Visual results on the Sth-Else dataset

Fig. 8 shows a few examples of action recognition results of our **CDTD** on the Sth-Else dataset.

Figure 8: The visual examples of **CDTD**$_{NCE}$ on Sth-Else dataset. **CDTD**$_{NCE}$ correctly predicts the four actions but misclassifies two videos of "lifting a surface with something on it", and "Burying something in something" as "Moving a part of something"

D. Additional Experiments

In this section, we focus on comparisons with state-of-the-art metric-based methods, including MoLo (Wang et al., 2023b), HySRM (Wang et al., 2022), HCL (Zheng et al., 2022), OTAM (Cao et al., 2020), TRX (Perrett et al., 2021a) and STRM (Thatipelli et al., 2022). For a fair evaluation, we perform experiments across the SSv2, UCF-101, HMDB-51, and Kinetics datasets. In the SSv2-Full and SSv2-Small datasets, we randomly selected 64 classes for $\mathcal{D}$ and 24 for $\mathcal{S}$ and $\mathcal{Q}$. The main difference between SSv2-Full and SSv2-Small is the dataset size, with SSv2-Full containing all samples per category and SSv2-Small including only 100 samples per category. For HMDB-51, we chose 31 action categories for $\mathcal{D}$ and 10 for $\mathcal{S}$ and $\mathcal{Q}$, while for UCF-101, the selection was 70 and 21 categories, respectively. For Kinect, we used 64 action categories for $\mathcal{D}$ and 24 for $\mathcal{S}$ and $\mathcal{Q}$. To maintain statistical significance, we executed 200 trials, each involving random samplings across categories. After training on $\mathcal{D}$, we used k video sequences from each action category to form $\mathcal{S}$ for model updates. The inference phase utilized the remaining data from $\mathcal{Q}$.

Table 6 and Table 7 show additional experiments for 5-way-k-shot learning where the base and novel data are taken from the same original dataset. Between the two tables we observe that **CDTD**$_{CE}$ achieves the highest Top-1 accuracy in 11 out of 12 experiments, coming in second by only 0.4 with $k = 5$ on the UCF-101 dataset.

Table 6: Comparing **CDTD**$_{CE}$ to benchmarks for 5-way-k-shot learning on the SSv2 and SSv2-small using the ResNet-50 backbone

	SSv2			SSv2-small		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
OTAM	42.8	51.5	52.3	36.4	45.9	48.0
TRX	42.0	57.6	62.6	36.0	51.9	56.7
STRM	42.0	59.1	68.1	37.1	49.2	55.3
HyRSM	54.3	65.1	69.0	40.6	52.3	56.1
HCL	47.3	59.0	64.9	38.7	49.1	55.4
MoLo	<u>56.6</u>	<u>67.0</u>	<u>70.6</u>	<u>42.7</u>	<u>52.9</u>	<u>56.4</u>
CDTD$_{CE}$	**60.0**	**68.3**	**71.9**	**45.8**	**53.6**	**58.0**

Table 7: Comparing **CDTD**$_{CE}$ to benchmarks for 5-way-k-shot learning on the UCF-101, HMDB-51, and Kinectics datasets using the ResNet-50 backbone.

	UCF-101		HMDB-51		Kinects	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
OTAM	79.9	88.9	54.5	68.0	79.9	88.9
TRX	78.2	96.1	53.1	75.6	78.2	96.2
STRM	80.5	**96.9**	52.3	77.3	80.5	<u>96.9</u>
HyRSM	83.9	94.7	60.3	76.0	83.9	94.7
HCL	82.8	93.3	59.1	76.3	73.7	85.8
MoLo	<u>86.0</u>	95.5	<u>60.8</u>	<u>77.4</u>	<u>86.0</u>	95.5
CDTD$_{CE}$	**87.3**	<u>96.5</u>	**61.9**	**80.5**	**86.1**	**98.0**

D.1. Comparisons to Augmented Models

We further assess if adding $\mathbf{u}$ to other methods would further improve the results in Table 8. The results indicate marginal improvements over the original methods. Again, our **CDTD** obtains the highest accuracy among these methods.

E. Network Architectures

Tab. 9 illustrates the details of our implementation on **CDTD**$_{NCE}$.

Table 8: Additional comparisons by augmenting existing methods on the Sth-Else dataset using the ViT-B/16 backbone. $+\mathbf{u}$ means the method updates the domain encoder when adapting instead of fine-tuning. Since VL Prompting uses VPT (Jia et al., 2022) within the ViFi-CLIP framework, we only test ViFi-CLIP$+\mathbf{u}$.

	Sth-Else	
	5-shot	10-shot
ORViT	33.3	40.2
ORViT + $\mathbf{u}$	33.9	41.8
SViT	34.4	<u>42.6</u>
SViT + $\mathbf{u}$	<u>35.2</u>	**44.0**
CDTD$_{CE}$	**37.6**	**44.0**
ViFi-CLIP	44.5	54.0
VL Prompting	44.9	<u>58.2</u>
ViFi-CLIP + $\mathbf{u}$	<u>45.2</u>	58.0
CDTD$_{NCE}$	**48.5**	**59.9**

Table 9: The details of our network architectures for **CDTD**$_{NCE}$, where BS means batch size.

Configuration	Description	Output dimensions
Image encoder		
Input: concat($\mathbf{x}_{1:T}$)		BS $\times T \times 1024$
Dense	256 neurons, LeakyReLU	BS $\times T \times 256$
Dense	256 neurons, LeakyReLU	BS $\times T \times 256$
Dense	Temporal embeddings	BS $\times T \times 2N$
Bottleneck	Compute mean and variance of posterior	μ, σ
Reparameterization	Sequential sampling	$\hat{\mathbf{z}}_{1:T}$
Temporal dynamic generation		
Input: $\hat{\mathbf{z}}_{1:T}$		BS $\times T \times N$
Dense	256 neurons, LeakyReLU	BS $\times T \times 256$
Dense	256 neurons, LeakyReLU	BS $\times T \times 256$
Dense	input embeddings	BS $\times T \times 1024$
Temporal dynamic transition		
Input	$\hat{\mathbf{z}}_{1:T}$	BS $\times T \times N$
InverseTransition	ϵ_t	BS $\times T \times N$
JacobianCompute	$\log\det\lvert J\rvert$	BS
Classifier		
Input: Concat($\hat{\mathbf{z}}_{1:T}$)		BS $\times T \times N$
Dense	256 neurons, LeakyReLU	BS $\times T \times 256$
Dense	256 neurons, LeakyReLU	BS $\times T \times 256$
Dense	output embeddings	BS $\times T \times 1024$