Learning Calibratable Policies using Programmatic Style-Consistency

Eric Zhan 1  Albert Tseng 1  Yisong Yue 1  Adith Swaminathan 2  Matthew Hausknecht 2

Abstract
We study the problem of controllable generation of long-term sequential behaviors, where the goal is to calibrate to multiple behavior styles simultaneously. In contrast to the well-studied areas of controllable generation of images, text, and speech, there are two questions that pose significant challenges when generating long-term behaviors: how should we specify the factors of variation to control, and how can we ensure that the generated behavior faithfully demonstrates combinatorially many styles? We leverage programmatic labeling functions to specify controllable styles, and derive a formal notion of style-consistency as a learning objective, which can then be solved using conventional policy learning approaches. We evaluate our framework using demonstrations from professional basketball players and agents in the MuJoCo physics environment, and show that existing approaches that do not explicitly enforce style-consistency fail to generate diverse behaviors whereas our learned policies can be calibrated for up to $4^5 (1024)$ distinct style combinations.

1. Introduction
The widespread availability of recorded tracking data is enabling the study of complex behaviors in many domains, including sports (Chen et al., 2016a; Le et al., 2017b; Zhan et al., 2019; Yeh et al., 2019), video games (Kurin et al., 2017; Broll et al., 2019; Hofmann, 2019), laboratory animals (Eyjolfsdottir et al., 2014; 2017; Branson et al., 2009; Johnson et al., 2016), facial expressions (Suwajanakorn et al., 2017; Taylor et al., 2017), commonplace activities such as cooking (Nishimura et al., 2019), and transportation (Bojarski et al., 2016; Luo et al., 2018; Li et al., 2018; Chang et al., 2019). A key aspect of modern behavioral datasets is that the behaviors can exhibit very diverse styles (e.g., from multiple demonstrators). For example, Figure 1a depicts demonstrations from basketball players with variations in speed, desired destinations, and curvature of movement.

The goal of this paper is to study controllable generation of diverse behaviors by learning to imitate raw demonstrations; or more technically, to develop style-calibrated imitation learning methods. A controllable, or calibratable, policy would enable the generation of behaviors consistent with various styles, such as low movement speed (Figure 1b), or approaching the basket (Figure 1c), or both styles simultaneously (Figure 1d). Style-calibrated imitation learning methods that can yield such policies can be broadly useful to: (a) perform more robust imitation learning from diverse demonstrations (Wang et al., 2017; Broll et al., 2019), (b) enable diverse exploration in reinforcement learning agents (Co-Reyes et al., 2018), or (c) visualize and extrapolate counterfactual behaviors beyond those seen in the dataset (Le et al., 2017a), amongst many other tasks.

Performing style-calibrated imitation is a challenging task. First, what constitutes a “style”? Second, when can we be certain that a policy is “calibrated” when imitating a style? Third, how can we scale policy learning to faithfully generate combinatorially many styles? In related tasks like controllable image generation, common approaches for calibration use adversarial information factorization or mutual information between generated images and user-specified styles (e.g. gender, hair length, etc.) (Creswell et al., 2017; Lample et al., 2017; Chen et al., 2016b). However, we find that these indirect approaches fall well short of generating calibratable sequential behaviors. Intuitively, the aforementioned objectives provide only indirect proxies for style-calibration. For example, Figure 2 illustrates that an indirect baseline approach struggles to reliably generate trajectories to reach a certain displacement, even though the dataset contains many examples of such behavior.

Research questions. We seek to answer three research questions while tackling this challenge. The first is strategic: since high-level stylistic attributes like movement speed are typically not provided with the raw demonstration data, what systematic form of domain knowledge can we leverage to quickly and cleanly extract highly varied style information from raw behavioral data? The second is formulaic: how can
Learning Calibratable Policies using Programmatic Style-Consistency

(a) Expert demonstrations  (b) Style: SPEED  (c) Style: DESTINATION  (d) Both styles
Figure 1. Basketball trajectories from policies that are: (a) the expert; (b) calibrated to move at low speeds; (c) calibrated to end near the basket (within green boundary); and (d) calibrated for both (b,c) simultaneously. Diamonds (♦) and dots (●) are initial and final positions.

(a) Baseline, low displacement  (b) Ours, low displacement  (c) Baseline, high displacement  (d) Ours, high displacement
Figure 2. Basketball trajectories sampled from baseline policies and our models calibrated to the style of DISPLACEMENT with 6 classes corresponding to regions separated by blue lines. Diamonds (♦) and dots (●) indicate initial and final positions respectively. Each policy is conditioned on a label class for DISPLACEMENT (low in (a,b), high in (c,d)). Green dots indicate trajectories that are consistent with the style label, while red dots indicate those that are not. Our policy (b,d) is better calibrated for this style than the baselines (a,c).

we formalize the learning objective to encourage learning style-calibratable policies that can be controlled to realize many diverse styles? The third is algorithmic: how do we design practical learning approaches that reliably optimize the learning objective?

Our contributions. To address these questions, we present a novel framework inspired by data programming (Ratner et al., 2016), a paradigm in weak supervision that utilizes automated labeling procedures, called labeling functions, to learn without ground-truth labels. In our setting, labeling functions enable domain experts to quickly translate domain knowledge of diverse styles into programmatically generated style annotations. For instance, it is trivial to write programmatic labeling functions for the styles depicted in Figures 1 & 2 (speed and destination). Labeling functions also motivate a new learning objective, which we call programmatic style-consistency: rollouts generated by a policy calibrated for a particular style should return the same style label when fed to the programmatic labeling function. This notion of style-consistency provides a direct approach to measuring how calibrated a policy is, and does not suffer from the weaknesses of indirect approaches such as mutual information estimation. In the basketball example of scoring when near the basket, trajectories that perform correlated events (like turning towards the basket) will not return the desired style label when fed to the labeling function that checks for scoring events. We elaborate on this in Section 4.

We demonstrate style-calibrated policy learning in Basketball and MuJoCo domains. Our experiments highlight the modularity of our approach – we can plug-in any policy class and any imitation learning algorithm and reliably optimize for style-consistency using the approach of Section 5. The resulting learned policies can achieve very fine-grained and diverse style-calibration with negligible degradation in imitation quality – for example, our learned policy is calibrated to $4^5(1024)$ distinct style combinations in Basketball.

2. Related Work

Our work combines ideas from policy learning and data programming to develop a weakly supervised approach for more explicit and fine-grained calibration. As such, our work is related to learning disentangled representations and controllable generative modeling, reviewed below.

Imitation learning of diverse behaviors has focused on unsupervised approaches to infer latent variables/codes that
capture behavior styles (Li et al., 2017; Hausman et al., 2017; Wang et al., 2017). Similar approaches have also been studied for generating text conditioned on attributes such as sentiment or tense (Hu et al., 2017). A typical strategy is to maximize the mutual information between the latent codes and trajectories, in contrast to our notion of programmatic style-consistency.

**Disentangled representation learning** aims to learn representations where each latent dimension corresponds to exactly one desired factor of variation (Bengio et al., 2012). Recent studies (Locatello et al., 2019) have noted that popular techniques (Chen et al., 2016b; Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018) can be sensitive to hyperparameters and that evaluation metrics can be correlated with certain model classes and datasets, which suggests that fully unsupervised learning approaches may, in general, be unreliable for discovering cleanly calibratable representations. We avoid this roadblock by relying on programmatic labeling functions to provide weak supervision.

**Conditional generation** for images has recently focused on attribute manipulation (Bao et al., 2017; Creswell et al., 2017; Klys et al., 2018), which aims to enforce that changing a label affects only one aspect of the image (similar to disentangled representation learning). We extend these models and compare with our approach in Section 6. Our experiments suggest that these algorithms do not necessarily scale well into sequential domains.

**Enforcing consistency in generative modeling**, such as cycle-consistency in image generation (Zhu et al., 2017), and self-consistency in hierarchical reinforcement learning (Co-Reyes et al., 2018) has proved beneficial. The former minimizes a discriminative disagreement, whereas the latter minimizes a distributional disagreement between two sets of generated behaviors (e.g., KL-divergence). From this perspective, our style-consistency notion is more similar to the former: however we also enforce consistency over multiple time-steps, which is more similar to the latter.

**Goal-conditioned policy learning** considers policies that take as input the current state along with a desired goal state (e.g., a location), and then must execute a sequence of actions to achieve the goal states. In some cases, the goal states are provided exogenously (Zheng et al., 2016; Le et al., 2018; Broll et al., 2019; Ding et al., 2019), and in other cases the goal states are learned as part of a hierarchical policy learning approach (Co-Reyes et al., 2018; Sharma et al., 2020) in a way that uses a self-consistency metric similar to our style-consistency approach. Our approach can be viewed as complementary to these approaches as the goal is to study more general notions of consistency (e.g., our styles subsume goals as a special case) as well as to scale to combinatorial joint style spaces.

Hierarchical control via learning latent motor dynamics is concerned with recovering a latent representation of motor control dynamics such that one can easily design controllers in the latent space (which then get decoded into actions). The high level controllers can then be designed afterwards in a pipelined workflow (Lossy et al., 2020; Ling et al., 2020; Luo et al., 2020). The controllers are effective for short time horizons and focus on finding good representations of complex dynamics, whereas we focus on controlling behavior styles that can span longer horizons.

### 3. Background: Imitation Learning for Behavior Trajectories

Since our focus is on learning style-calibratable generative policies, for simplicity we develop our approach with the basic imitation learning paradigm of behavioral cloning. Interesting future directions include composing our approach with more advanced imitation learning approaches like DAGGER (Ross et al., 2011), GAIL (Ho & Ermon, 2016) as well as with reinforcement learning.

**Notation.** Let $S$ and $A$ denote the environment state and action spaces. At each timestep $t$, an agent observes state $s_t \in S$ and executes action $a_t \in A$ using a policy $\pi : S \rightarrow A$. The environment then transitions to the next state $s_{t+1} \sim \tau$ according to a (typically unknown) dynamics function $f : S \times A \rightarrow S$. For the rest of this paper, we assume $f$ is deterministic; a modification of our approach for stochastic $f$ is included in Appendix B. A trajectory $\tau$ is a sequence of $T$ state-action pairs and the last state: $\tau = \{(s_t, a_t)\}_{t=1}^T \cup \{s_{T+1}\}$. Let $D$ be a set of $N$ trajectories collected from expert demonstrations. In our experiments, each trajectory in $D$ has the same length $T$, but in general this does not need to be the case.

**Learning objective.** We begin with the basic imitation learning paradigm of behavioral cloning (Syed & Schapire, 2008). The goal is to learn a policy that behaves like the pre-collected demonstrations:

$$\pi^* = \arg\min_{\pi} \mathbb{E}_{\tau \sim D} [L_{imitation}(\tau, \pi)],$$

where $L_{imitation}$ is a loss function that quantifies the mismatch between actions chosen by $\pi$ and those in the demonstrations. Since we are primarily interested in probabilistic or generative policies, we typically use (variants of) negative log-density: $L(\tau, \pi) = \sum_{t=1}^T - \log \pi(a_t|s_t)$, where $\pi(a_t|s_t)$ is the probability of $\pi$ picking action $a_t$ in $s_t$.

**Policy class of $\pi$.** Common model choices for instantiating $\pi$ include sequential generative models like recurrent Neural Networks (RNN) and trajectory variational autoencoders (TVAE). TVAEs introduce a latent variable $\mathbf{z}$ (also called a trajectory embedding), an encoder network $q_{\phi}$, a policy
Learning Calibratable Policies using Programmatic Style-Consistency

decoder $\pi_\theta$, and a prior distribution $p$ on $z$. They have been shown to work well in a range of generative policy learning settings (Wang et al., 2017; Ha & Eck, 2018; Co-Reyes et al., 2018), and have the following imitation learning objective:

$$L_{\text{tva}}(\tau, \pi_\theta; q_\phi) = \mathbb{E}_{q_\phi(z|\tau)} \left[ \sum_{t=1}^{T} - \log \pi_\theta(a_t|s_t, z) \right] + D_{KL}(q_\phi(z|\tau)||p(z)). \quad (2)$$

The first term in (2) is the standard negative log-density that the policy assigns to trajectories in the dataset, while the second term is the KL-divergence between the prior and approximate posterior of trajectory embeddings $z$. The main shortcoming of VAEs and related approaches, which we address in Sections 4 & 5, is that the resulting policies cannot be easily calibrated to generate specific styles. For instance, the goal of the trajectory embedding $z$ is to capture all the styles that exist in the expert demonstrations, but there is no guarantee that the embeddings easily encode the desired styles in a calibrated way. Previous work has largely relied on unsupervised learning techniques that either require significant domain knowledge (Le et al., 2017b), or have trouble scaling to complex styles commonly found in real-world applications (Wang et al., 2017; Li et al., 2017).

4. Programmatic Style-consistency

Building upon the basic setup in Section 3, we focus on the setting where the demonstrations $D$ contain diverse behavior styles. To start, let $y \in Y$ denote a single style label (e.g., speed or destination, as shown in Figure 1). Our goal is to learn a policy $\pi$ that can be explicitly calibrated to $y$, i.e., trajectories generated by $\pi(\cdot|y)$ should match the demonstrations in $D$ that exhibit style $y$.

Obtaining style labels can be expensive using conventional annotation methods, and unreliable using unsupervised approaches. We instead utilize easily programmable labeling functions that automatically produce style labels. We then formalize a notion of style-consistency as a learning objective, and in Section 5 describe a practical learning approach.

Labeling functions. Introduced in the data programming paradigm (Ratner et al., 2016), labeling functions programatically produce weak and noisy labels to learn models on otherwise unlabeled datasets. A significant benefit is that labeling functions are often simple scripts that can be quickly applied to the dataset, which is much cheaper than manual annotations and more reliable than unsupervised methods. In our framework, we study behavior styles that can be represented as labeling functions, which we denote $\lambda$, that map trajectories $\tau$ to style labels $y$. For example:

$$\lambda(\tau) = 1\{|s_{t+1} - s_t|_2 > c\}, \quad (3)$$

which distinguishes between trajectories with large (greater than a threshold $c$) versus small total displacement. We experiment with a range of labeling functions, as described in Section 6. Many behavior styles used in previous work can be represented as labeling functions, e.g., agent speed (Wang et al., 2017). Multiple labeling functions can be provided at once resulting in a combinatorial space of joint style labels. We use trajectory-level labels $\lambda(\tau)$ in our experiments, but in general labeling functions can be applied on subsequences $\lambda(\tau_{t:t+h})$ to obtain per-timestep labels, e.g., agent goal (Broll et al., 2019). We can efficiently annotate datasets using labeling functions, which we denote as $\lambda(D) = \{(\tau_i, \lambda(\tau_i))\}_{i=1}^N$. Our goal can now be phrased as: given $\lambda(D)$, train a policy $\pi : S \times Y \mapsto A$ such that $\pi(\cdot|y)$ is calibrated to styles $y$ found in $\lambda(D)$.

Style-consistency. A key insight in our work is that labeling functions naturally induce a metric for calibration. If a policy $\pi(\cdot|y)$ is calibrated to $\lambda$, we would expect the generated behaviors to be consistent with the label. So, we expect the following loss to be small:

$$\mathbb{E}_{y \sim p(y), \tau \sim \pi(\cdot|y)} \left[ L_{\text{style}}(\lambda(\tau), y) \right], \quad (4)$$

where $p(y)$ is a prior over the style labels, and $\tau$ is obtained by executing the style-conditioned policy in the environment. $L_{\text{style}}$ is thus a disagreement loss over labels that is minimized at $\lambda(\tau) = y$, e.g., $L_{\text{style}}(\lambda(\tau), y) = 1 \{ \lambda(\tau) \neq y \}$ for categorical labels. We refer to (4) as the style-consistency loss, and say that $\pi(\cdot|y)$ is maximally calibrated to $\lambda$ when (4) is minimized. Our learning objective adds (1) with (4):

$$\pi^* = \arg \min_\pi \mathbb{E}_{y \sim p(y), \tau \sim \pi(\cdot|\lambda(D))} \left[ L_{\text{imitation}}(\tau, \pi(\cdot|\lambda(D))) \right] + \mathbb{E}_{y \sim p(y), \tau \sim \pi(\cdot|y)} \left[ L_{\text{style}}(\lambda(\tau), y) \right]. \quad (5)$$

The simplest choice for the prior distribution $p(y)$ is the marginal distribution of styles in $\lambda(D)$. The first term in (5) is a standard imitation learning objective and can be tractably estimated using $\lambda(D)$. To enforce style-consistency with the second term, conceptually we need to sample several $y \sim p(y)$, then several rollouts $\tau \sim \pi(\cdot|y)$ from the current policy, and query the labeling function for each of them. Furthermore, if $\lambda$ is a non-differentiable function defined over the entire trajectory, as is the case in (3), then we cannot simply backpropagate the style-consistency loss. In Section 5, we introduce differentiable approximations to more easily optimize the objective in (5).

Combinatorial joint style space. Our notion of style-consistency can be easily extended to optimize for combinatorially-many joint styles when multiple labeling functions are provided. Suppose we have $M$ labeling functions $\{\lambda_i\}_{i=1}^M$ and corresponding label spaces $\{Y_i\}_{i=1}^M$. Let $\lambda$ denote $(\lambda_1, \ldots, \lambda_M)$ and $y$ denote $(y_1, \ldots, y_M)$. Style-
Learning Calibratable Policies using Programmatic Style-Consistency

consistency loss becomes:

\[ \mathbb{E}_{y \sim p(y), \tau \sim \pi(\cdot|y)} \left[ \sum_{i=1}^{M} \mathcal{L}^\text{style}_i(\lambda_i(\tau), y_i) \right]. \tag{6} \]

Note that style-consistency is optimal when the generated trajectory agrees with all labeling functions. Although challenging to achieve, this outcome is most desirable, i.e., \( \pi(\cdot|y) \) is calibrated to all styles simultaneously. Indeed, a key metric that we evaluate is how well various learned policies can be calibrated to all styles simultaneously (i.e., loss of 0 only if all styles are calibrated, and loss of 1 otherwise).

5. Learning Approach

Optimizing (5) is challenging due to the long-time horizon and non-differentiability of the labeling functions \( \lambda \). Given unlimited queries to the environment, one could naively employ model-free reinforcement learning, e.g., estimating (4) using rollouts and optimizing using policy gradient approaches. We instead take a model-based approach, described generically in Algorithm 1, that is more computationally-efficient and decomposable (i.e., transparent). The model-based approach is compatible with batch offline learning, and we found it particularly useful for diagnosing deficiencies in our algorithmic framework. We first introduce a label approximator \( \lambda \) and then show how to optimize through the environmental dynamics using a differentiable model-based learning approach.

Approximating labeling functions. To deal with non-differentiability of \( \lambda \), we approximate it with a differentiable function \( \psi^* \) parameterized by \( \psi \):

\[ \psi^* = \arg \min_{\psi} \mathbb{E}_{(\tau, \lambda(\tau)) \sim \lambda(D)} \left[ \mathcal{L}^\text{label}(C^\lambda_{\psi}(\tau), \lambda(\tau)) \right]. \tag{7} \]

Here, \( \mathcal{L}^\text{label} \) is a differentiable loss that approximates \( \mathcal{L}^\text{style} \), such as cross-entropy loss when \( \mathcal{L}^\text{style} \) is the 0/1 loss. In our experiments we use a RNN to represent \( C^\lambda_{\psi} \). We then modify the style-consistency term in (5) with \( C^\lambda_{\psi} \). and optimize:

\[ \pi^* = \arg \min_{\pi} \mathbb{E}_{(\tau, \lambda(\tau)) \sim \lambda(D)} \left[ \mathcal{L}^\text{imitation}(\tau, \pi(\cdot|\lambda(\tau))) \right] + \mathbb{E}_{y \sim p(y), \tau \sim \pi(\cdot|y)} \left[ \mathcal{L}^\text{label}(C^\lambda_{\psi^*}(\tau), y) \right]. \tag{8} \]

Optimizing \( \mathcal{L}^\text{style} \) over trajectories. The next challenge is to optimize style-consistency over multiple time steps. Consider the labeling function in (3) that computes the difference between the first and last states. Our label approximator \( C^\lambda_{\psi^*} \) may converge to a solution that ignores all inputs except for \( s_1 \) and \( s_{T+1} \). In this case, \( C^\lambda_{\psi^*} \) provides no learning signal about intermediate steps. As such, effective optimization of style-consistency in (8) requires informative learning signals on all actions at every step, which can be viewed as a type of credit assignment problem.

In general, model-free and model-based approaches address this challenge in dramatically different ways and for different problem settings. A model-free solution views this credit assignment challenge as analogous to that faced by reinforcement learning (RL), and repurposes generic reinforcement learning algorithms. Crucially, they assume access to the environment to collect more rollouts under any new policy. A model-based solution does not assume such access and can operate only with the batch of behavior data \( D \); however they can have an additional failure mode since the learned models may provide an inaccurate signal for proper credit assignment. We choose a model-based approach, which exploits access to the environment when available to refine the learned models, for two reasons: (a) we found it to be compositionally simpler and easier to debug; and (b) we can use the learned model to obtain hallucinated rollouts of any policy efficiently during training.

Modeling dynamics for credit assignment. Our model-based approach utilizes a dynamics model \( M \) to approximate the environment’s dynamics by predicting the change in state given the current state and action:

\[ \varphi^* = \arg \min_{\varphi} \mathbb{E}_{\tau \sim D} \sum_{t=1}^{T} \mathcal{L}^\text{dynamics}(M_{\varphi}(s_t, a_t), (s_{t+1} - s_t)), \tag{9} \]

where \( \mathcal{L}^\text{dynamics} \) is often \( L_2 \) or squared-\( L_2 \) loss (Nagabandi et al., 2018; Luo et al., 2019). This allows us to generate trajectories by rolling out: \( s_{t+1} = s_t + M_{\varphi}(s_t, \pi(s_t)) \). Then optimizing for style-consistency in (8) would backpropagate through our dynamics model \( M_{\varphi} \) and provide informative learning signals to the policy at every timestep.

We outline our model-based approach in Algorithm 2. Lines 12-15 describe an optional step to fine-tune the dynamics model by querying the environment using the current policy (similar to Luo et al. (2019)); we found that this can improve style-consistency in some experiments. In Appendix B we elaborate how the dynamics model and objective of Eqn (9) is changed if the environment is stochastic.

Discussion. To summarize, we claim that style-consistency
Algorithm 2 Model-based approach for Algorithm 1

1: Input: demonstrations $\mathcal{D}$, labeling function $\lambda$, label approximator $C_{\lambda}$, dynamics model $M_{\theta}$
2: $\lambda(\mathcal{D}) \leftarrow \{ (\tau_i, \lambda(\tau_i)) \}_{i=1}^{N}$
3: for $n_{\text{dynamics}}$ iterations do
4: optimize (9) with batch from $\mathcal{D}$
5: end for
6: for $n_{\text{label}}$ iterations do
7: optimize (7) with batch from $\lambda(\mathcal{D})$
8: end for
9: for $n_{\text{policy}}$ iterations do
10: $B \leftarrow \{ n_{\text{collect}} \text{ trajectories using } M_{\theta} \text{ and } \pi \}$
11: optimize (8) with $\lambda(\mathcal{D})$ and $B$
12: for $n_{\text{env}}$ iterations do
13: $\tau_{\text{env}} \leftarrow \{ 1 \text{ trajectory using environment and } \pi \}$
14: optimize (9) with $\tau_{\text{env}}$
15: end for
16: end for

is an “objective” metric to measure the quality of calibration. Our learning approach uses off-the-shelf methods to enforce style-consistency during training. We anticipate several variants of style-consistent policy learning of Algorithm 1 — for example, using model-free RL, using environment/model rollouts to fine-tune the labeling function approximator, using style-conditioned policy classes, or using other loss functions to encourage imitation quality. Our experiments in Section 6 establish that our style-consistency loss provides a clear learning signal, that no prior approach directly enforces this consistency, and that our approach accomplishes calibration for a combinatorial joint style space.

6. Experiments

We first briefly describe our experimental setup and baseline choices, and then discuss our main experimental results. A full description of experiments is available in Appendix C.2

Data. We validate our framework on two datasets: 1) a collection of professional basketball player trajectories with the goal of learning a policy that generates realistic player-movement, and 2) a Cheetah agent running horizontally in MuJoCo (Todorov et al., 2012) with the goal of learning a policy with calibrated gaits. The former has a known dynamics function: $f(s_t, a_t) = s_t + a_t$, where $s_t$ and $a_t$ are the player’s position and velocity on the court respectively; we expect the dynamics model $M_{\theta}$ to easily recover this function. The latter has an unknown dynamics function (which we learn a model of when approximating style-consistency). We obtain Cheetah demonstrations from a collection of policies trained using pytorch-a2c-ppo-acktr (Kostrikov, 2018) to interface with the DeepMind Control Suite’s Cheetah domain (Tassa et al., 2018)—see Appendix C for details.

Labeling functions. Labeling functions for Basketball include: 1) average SPEED of the player, 2) DISPLACEMENT from initial to final position, 3) distance from final position to a fixed DESTINATION on the court (e.g. the basket), 4) mean DIRECTION of travel, and 5) CURVATURE of the trajectory, which measures the player’s propensity to change directions. For Cheetah, we have labeling functions for the agent’s 1) SPEED, 2) TORSO HEIGHT, 3) BACK-FOOT HEIGHT, and 4) FRONT-FOOT HEIGHT that can be trivially computed from trajectories extracted from the environment.

We threshold the aforementioned labeling functions into categorical labels (leaving real-valued labels for future work) and use (4) for style-consistency with $L^\text{style}$ as the 0/1 loss. We use cross-entropy for $L^\text{label}$ and list all other hyperparameters in Appendix C.

Metrics. We will primarily study two properties of the learned models in our experiments — imitation quality, and style-calibration quality. For measuring imitation quality of generative models, we report the negative log-density term in (2), also known as the reconstruction loss term in VAE literature (Kingma & Welling, 2014; Ha & Eck, 2018), which corresponds to how well the policy can reconstruct trajectories from the dataset.

To measure style-calibration, we report style-consistency results as $1 - L^\text{style}$ in (4) so that all results are easily interpreted as accuracies. In Section 6.5, we find that style-consistency indeed captures a reasonable notion of calibration — when the labeling function is inherently noisy and style-calibration is hard, style-consistency correspondingly decreases. In Section 6.3, we find that the goals of imitation (as measured by negative log-density) and calibration (as measured by style-consistency) may not always be aligned — investigating this trade-off is an avenue for future work.

Baselines. Our main experiments use TVAEs as the underlying policy class. In Section 6.4, we also experiment with an RNN policy class. We compare our approach, CTVAE-style, with 3 baselines:

1. CTVAE: conditional TVAEs (Wang et al., 2017).
2. CTVAE-info: CTVAE with information factorization (Creswell et al., 2017), indirectly maximizes style-consistency by removing correlation of $y$ with $z$.
3. CTVAE-mi: CTVAE with mutual information maximization between style labels and trajectories. This is a supervised variant of unsupervised models (Chen et al., 2016b; Li et al., 2017), and also requires learning a dynamics model for sampling policy rollouts.

2Code is available at: https://github.com/ezhan94/calibratable-style-consistency.
Detailed descriptions of baselines are in Appendix A. All baseline models build upon TVAEs, which are also conditioned on a latent variable (see Section 3) and only fundamentally differ in how they encourage the calibration of policies to different style labels. We highlight that the underlying model choice is orthogonal to our contributions; our framework is compatible with other policy models (see Section 6.4).

**Model details.** We model all trajectory embeddings $z$ as a diagonal Gaussian with a standard normal prior. Encoder $q_\phi$ and label approximators $C_\psi^a$ are bi-directional GRUs (Cho et al., 2014) followed by linear layers. Policy $\pi_\theta$ is recurrent for basketball, but non-recurrent for Cheetah. The Gaussian log sigma returned by $\pi_\theta$ is state-dependent for basketball, but state-independent for Cheetah. For Cheetah, we made these choices based on prior work in MuJoCo for training gait policies (Wang et al., 2017). For Basketball, we observed a lot more variation in the 500k demonstrations so we experimented with a more flexible model. See Appendix C for hyperparameters.

### 6.1. How well can we calibrate policies for single styles?

We first threshold labeling functions into 3 classes for Basketball and 2 classes for Cheetah; the marginal distribution $p(y)$ of styles in $\lambda(D)$ is roughly uniform over these classes. Then we learn a policy $\pi^*$ calibrated to each of these styles. Finally, we generate rollouts from each of the learned policies to measure style-consistency. Table 1 compares the median style-consistency (over 5 seeds) of policies evaluated with 4,000 Basketball and 500 Cheetah rollouts. Trained separately for each style, CTVAE-style policies outperform baselines for all styles in Basketball and Cheetah environments.

![Table 1. Individual Style Calibration](image)

CTVAE-style degrades as the number of classes increases, it outperforms baselines for all styles in Basketball and Cheetah environments.

### 6.2. Can we calibrate for combinatorial joint style spaces?

We now consider combinatorial style-consistency as in (6), which measures the style-consistency with respect to all labeling functions simultaneously. For instance, in Figure 3, we calibrate to both terminating close to the net and also the speed at which the agent moves towards the target destination; if either style is not calibrated then the joint style is not calibrated. In our experiments, we evaluated up to 1024 joint styles.

Table 2 compares the style-consistency of policies simultaneously calibrated for up to 5 labeling functions for Basketball and 3 labeling functions for Cheetah. This is a very difficult task, and we see that style-consistency for base-
style-consistency without any degradation in imitation quality, since we optimize both in (5). For Basketball, high style-consistency is achieved without any degradation in imitation quality. However, we note that the labeling functions used thus far are assumed to be perfect, in that they capture exactly the style that we wish to calibrate. In practice, domain experts may specify labeling functions that are noisy; we simulate that scenario in this experiment.

In Table 3, we investigate whether CTVAE-style’s superior style-consistency comes at a significant cost to imitation quality. We visualize a CTVAE AE-style policy calibrated for two styles in Figure 3. CTVAE-style rollouts calibrated for 2 styles: label class 1 of DESTINATION (net) (see Figure 5 in Appendix D) and each class for SPEED, with 0.93 style-consistency. Diamonds (●) and dots (○) indicate initial and final positions.

Table 3. Combinatorial Style-consistency: \( \times 10^{-2} \), median over 5 seeds) Simultaneously calibrated to joint styles from multiple labeling functions, CTVAE-style policies significantly outperform all baselines. The number of distinct style combinations are in brackets. The most challenging experiment for basketball calibrates for 1024 joint styles (5 labeling functions, 4 classes each), in which CTVAE-style has a +161% improvement in style-consistency over the best baseline.

Table 4. KL-divergence and negative log-density per timestep for TVAE models (lower is better). CTVAE-style is comparable to baselines for Basketball, but is slightly worse for Cheetah.

Table 5. Style-consistency of RNN policy model (\( 10^{-2} \), 5 seeds) for DESTINATION in basketball. Our approach improves style-consistency without significantly decreasing imitation quality. For Cheetah, negative log-density is slightly worse; a followup experiment in Table 13 in Appendix D shows that we can improve imitation quality with more training, sometimes with modest decrease to style-consistency.

6.4. Is our framework compatible with other policy classes for imitation?

We highlight that our framework introduced in Section 5 is compatible with any policy class. In this experiment, we optimize for style-consistency using a simpler model for the policy and show that style-consistency is still improved. In particular, we use an RNN and calibrate for DESTINATION in basketball. In Table 5, we see that style-consistency is improved for the RNN model without any significant decrease in imitation quality.

6.5. What if labeling functions are noisy?

So far, we have demonstrated that our method optimizing for style-consistency directly can learn policies that are much better calibrated to styles, without a significant degradation in imitation quality. However, we note that the labeling functions used thus far are assumed to be perfect, in that they capture exactly the style that we wish to calibrate. In practice, domain experts may specify labeling functions that are noisy; we simulate that scenario in this experiment.
We believe that our framework lays the foundation for many directions of future research. First, can one model more complex styles not easily captured with a single labeling function (e.g. aggressive vs. passive play in sports) by composing simpler labeling functions (e.g. max speed, distance to closest opponent, number of fouls committed, etc.), similar to (Ratner et al., 2016; Bach et al., 2017)? Second, can we use these per-timestep labels to model transient styles, or simplify the credit assignment problem when learning to calibrate? Third, can we blend our programmatic supervision with unsupervised learning approaches to arrive at effective semi-supervised solutions? Fourth, can we use model-free approaches to further optimize self-consistency, e.g., to fine-tune from our model-based approach? Finally, can we integrate our framework with reinforcement learning to also optimize for environmental rewards?

**Acknowledgements**

This research is supported in part by NSF #1564330, NSF #1918655, DARPA PAI, and gifts from Intel, Activision/Blizzard and Northrop Grumman. Basketball dataset was provided by STATS.

**References**


Learning Calibratable Policies using Programmatic Style-Consistency


Learning Calibratable Policies using Programmatic Style-Consistency


