2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Beyond Domain Randomization: Event-Inspired Perception for Visually Robust Adversarial Imitation from Videos

Andrea Ramazzina*[1,2], Vittorio Giammarino*[3], Matteo El-Hariry[4] and Mario Bijelic[5,6]

Abstract— **Imitation from videos often fails when expert demonstrations and learner environments exhibit domain shifts, such as discrepancies in lighting, color, or texture. While visual randomization partially addresses this problem by augmenting training data, it remains computationally intensive and inherently reactive, struggling with unseen scenarios. We propose a different approach: instead of randomizing appearances, we eliminate their influence entirely by rethinking the sensory representation itself. Inspired by biological vision systems that prioritize temporal transients (e.g., retinal ganglion cells) and by recent sensor advancements, we introduce event-inspired perception for visually robust imitation. Our method converts standard RGB videos into a sparse, event-based representation that encodes temporal intensity gradients, discarding static appearance features. This biologically grounded approach disentangles motion dynamics from visual style, enabling robust visual imitation from observations even in the presence of visual mismatches between expert and agent environments. By training policies on event streams, we achieve invariance to appearance-based distractors without requiring computationally expensive and environment-specific data augmentation techniques. Experiments across the DeepMind Control Suite and the Adroit platform for dynamic dexterous manipulation show the efficacy of our method. Our code is publicly available at this link.**

I. INTRODUCTION

Visual Imitation from Observations (V-IfO) has emerged as a rapidly growing research area in recent years [1], [2]. This approach presents a compelling proposition for the future of robotics, offering an efficient and scalable method for teaching new skills to autonomous systems, and a possible alternative to tedious and complex handcrafting of ad-hoc rewards. While significant progress has been made in addressing the fundamental challenges of V-IfO—namely, the partial observability of decision-making processes and the lack of explicit expert action information—current solutions still face important limitations. Specifically, state-of-the-art (SOTA) end-to-end algorithms typically assume environmental consistency between expert demonstrations and agent deployment scenarios [1], [2]. This assumption proves problematic in real-world applications, where environmental conditions can vary significantly.

Recently, various methods [3]–[7] have been proposed to address this gap; with the most effective approach [7] relying on contrastive learning and data augmentation for invariant feature extraction. This reliance limits the applicability of this approach, as visual data augmentation techniques may be unavailable in certain domains or computationally prohibitive. Additionally, designing effective augmentations requires prior knowledge of the expert domain or assumptions on the types of mismatches, further restricting its usability.

We depart from this line of work and propose to eliminate the visual domain gap at its origin, namely the input signal representation. Specifically, we draw inspiration from biological vision systems, where retinal ganglion cells encode temporal intensity changes [8] rather than absolute luminance, and from the recent advancements in event cameras and their underlying image formation model. Building on these insights, we redefine the agent's perceptual interface to focus on motion-generated gradients. Consequently, our method inherently discards appearance-based distractors (lighting, color, texture) while preserving the dynamics features critical for control. We evaluate this technique, integrated with a SOTA end-to-end V-IfO algorithm, on two challenging benchmarks: the DeepMind Control Suite [9] for locomotion under visual perturbations (e.g., color randomization, lighting change) and the Adroit platform for robotic dexterous manipulation [10], requiring fine-grained motion alignment. Specifically, we make the following contributions:

- We design a lightweight event-inspired perception module that eliminates appearance-based distractors (e.g., color, texture, lighting) while preserving motion-critical temporal gradients.
- We provide a computationally efficient method that synthesizes event streams directly from RGB videos, bypassing computationally expensive domain randomization or handcrafted augmentation pipelines.
- By demonstrating robust imitation from synthetic event streams, we unlock the potential for real-world deployment with event cameras. Event cameras are low-power, high-temporal-resolution sensors that natively capture the temporal dynamics our method exploits.

II. RELATED WORK

a) Imitation from expert videos: The Imitation Learning (IL) paradigm involves training agents to replicate expert behavior using task demonstrations, typically represented as state-action pairs. Among IL methods, Adversarial Imitation Learning (AIL) [11], [12] stands out as a flexible and effective approach, modeling IL as an adversarial interaction between a discriminator and the agent's policy, where the discriminator distinguishes whether a state-action pair originates from the expert's or the agent's behavior policy. Building on principles from inverse Reinforcement Learning (RL) [13]–[16], AIL uses the discriminator's output as a reward signal to guide

*Denotes equal contribution

[1] Mercedes-Benz AG, Germany, [2] Technical University of Munich, Germany, [3] Purdue University, USA, [4] Space Robotics Research Group, SnT, University of Luxembourg, [5] Torc Robotics, USA, [6] Princeton University, USA, [7] https://github.com/VittorioGiammarino/Eb-LAIfO

the agent's training through RL. Recent adaptations of AIL extend its applicability to partially observable environments, addressing missing information [17], and to visual IL, where agents learn directly from video frames instead of structured states [18]. A key variation, Imitation from Observation [19]–[21], eliminates the need for action labels, making it more practical but also more challenging, as agents must infer expert behavior solely from state sequences. When these state sequences consist of video data without action labels, the problem is referred to as V-IfO. SOTA V-IfO methods include PatchAIL [1], which applies AIL directly to pixel space using a PatchGAN discriminator [22], [23], and LAIfO [2], which learns a latent representation of agent states. However, these methods assume that the expert and the learner operate within the same decision-making framework, an assumption rarely met in real-world applications.

b) Imitation from videos with mismatches: Our research tackles visual imitation from observations under domain mismatches (V-IfO with mismatches), a problem also known as third-person IL [3], domain-adaptive IL [24], or cross-domain IL [25]. Prior solutions fall into two categories: sequential and end-to-end approaches. Sequential methods break the problem into distinct learning stages, tackling them consecutively [26]–[30]. In contrast, end-to-end methods [3]–[7] aim to learn online, mapping pixel observations directly to actions without intermediate learning stages. In end-to-end methods, a common strategy for handling domain mismatches is to extract domain-invariant features. For instance, in [3], adversarial learning is used to align feature distributions across domains, while DisentanGAIL [5] enforces a mutual information constraint to preserve task-relevant information in the feature space. More recently, C-LAIfO [7] has leveraged contrastive learning for domain-invariant feature extraction, where these features are used across the entire AIL pipeline for both reward inference and policy learning. Unlike these methods, our approach avoids the need for domain-invariant feature learning. Instead, we disentangle motion dynamics from visual style at the observation level, effectively reducing the problem to standard V-IfO without mismatches. This eliminates the need for costly feature extraction steps while ensuring robust imitation performance. Other works address domain mismatches using generative approaches, such as learning domain-agnostic latent dynamics [4] or transforming expert videos to match the agent's domain via cycle-consistent adversarial networks [6]. However, these methods are computationally demanding and challenging to implement, as they rely on costly generative steps. In contrast, our approach filters out non-essential signals directly at the observation level, eliminating the need for additional learning steps. As a result, it is significantly more computationally efficient than learning-based alternatives.

c) Event-based cameras in robotics: Unlike conventional cameras that capture complete images at fixed intervals, event cameras detect brightness changes asynchronously and independently at each pixel. Event cameras produce a variable-rate stream of digital "events," with each event signaling a specific threshold change in brightness at a particular pixel location and time [31]. Specifically, each pixel stores the log intensity value whenever it generates an event and continuously monitors for significant deviations from this stored value [31]. Due to their high temporal resolution and low power characterisitics, event-based cameras have rapidly been adopted in vision-based robotics, enabling the development of highly responsive policies that directly convert sparse event data into control commands [32]. These systems have been deployed across various robotic platforms with promising results. For example, previous works have successfully trained quadruped robots to intercept fast-moving objects by applying RL to event-based visual inputs [33]. In aerial robotics, drones equipped with event cameras have demonstrated collision avoidance capabilities against dynamic obstacles using model-based algorithms [34], showcasing the exceptional temporal resolution of these sensors. The applications of event-based vision extend beyond robotics into computer vision domains, including classical computer vision tasks such as video deblurring [35]. Other lines of work have also developed multimodal approaches that integrate event-based and conventional frame-based vision, capitalizing on their complementary advantages. For instance, [36] proposed an end-to-end learned visual odometry framework that combines asynchronous event streams with standard image data, achieving improved efficiency and accuracy in challenging visual conditions.

While we do not directly use an event camera, we simulate its underlying functioning, extracting discrete events (in a synchronous manner) from a duplet of RGB frames, and use this representation as input for performing direct imitation from videos.

III. PRELIMINARIES

a) Modeling the visual mismatch in partially observable Markov decision process: In the standard formulation of the V-IfO problem, both the expert and the agent are assumed to operate within the same Partially Observable Markov Decision Process (POMDP), described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{X}, \mathcal{T}, \mathcal{U}, \mathcal{R}, \rho_0, \gamma)$. We define $\mathcal{S}$ as the state space, $\mathcal{A}$ the action space, $\mathcal{X}$ the observation space and $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$ the transition probability function where $P(\mathcal{S})$ denotes the set of probability distributions over $\mathcal{S}$. Furthermore, $\mathcal{U} : \mathcal{S} \to P(\mathcal{X})$ is the observation probability function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $\rho_0 \in P(\mathcal{S})$ the initial state distribution, and $\gamma \in [0, 1)$ the discount factor. In this work, we consider the expert and the agent interacting with different, but related, decision processes. Specifically, we define two distinct POMDPs: a *source-POMDP* for the expert and a *target-POMDP* for the agent. The target-POMDP is defined by $(\mathcal{S}, \mathcal{A}, \mathcal{X}, \mathcal{T}, \mathcal{U}_T, \mathcal{R}, \rho_0, \gamma)$, while the source-POMDP is characterized by $(\mathcal{S}, \mathcal{A}, \mathcal{X}, \mathcal{T}, \mathcal{U}_S, \mathcal{R}, \rho_0, \gamma)$. The key distinction between these two lies in their observation probability functions, $\mathcal{U}_S$ and $\mathcal{U}_T$, which govern the relationship between the environment state and the agent's or expert's observations. Specifically, the expert receives an observation $x_t^S \sim \mathcal{U}_S(\cdot|s_t)$ from the source-POMDP, whereas the agent observes $x_t^T \sim \mathcal{U}_T(\cdot|s_t)$ from the target-POMDP. Since these

observations may not be identical (i.e., $x_t^S \neq x_t^T$), we refer to this discrepancy as *visual mismatch*. This formulation captures scenarios where the agent must infer the expert's behavior despite differences in visual perception, a common challenge in real-world tasks such as sim-to-real transfer.

b) Reinforcement learning: Aligning with previous work, we define RL in the fully observable Markov Decision Process (MDP) which is a special case of a POMDP where the underlying state s is directly observable, i.e., $\mathcal{X} = \mathcal{S}$ and $\mathcal{U}(s) = \delta_s$, where δ_s denotes a Dirac distribution centered at s. Given an MDP and a stationary policy $\pi : \mathcal{S} \to P(\mathcal{A})$, the objective is to maximize the expected total discounted return, defined as $J(\pi) = \mathbb{E}_\tau[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)]$, where the trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$ is generated by following policy π. A stationary policy π induces a normalized discounted state visitation distribution given by $d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid \rho_0, \pi, \mathcal{T})$, representing the expected frequency of visiting state s under π. The corresponding normalized discounted state-action visitation distribution is $\rho_\pi(s,a) = d_\pi(s)\pi(a|s)$, which quantifies the expected frequency of encountering state-action pairs (s,a). Furthermore, we define the state-action value function (Q-function) as $Q^\pi(s,a) = \mathbb{E}_\tau[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t,a_t) \mid S_0 = s, A_0 = a]$. Finally, when a policy is parameterized by $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$, we denote it as $\pi_{\boldsymbol{\theta}}$.

c) Generative adversarial imitation learning: Assume we have a set of expert demonstrations $\tau_E = (s_{0:T}, a_{0:T})$ generated by the expert policy π_E, a set of trajectories $\tau_{\boldsymbol{\theta}}$ generated by the policy $\pi_{\boldsymbol{\theta}}$, and a discriminator network $D_{\boldsymbol{\chi}} : \mathcal{S} \times \mathcal{A} \to [0,1]$ parameterized by $\boldsymbol{\chi}$. Generative adversarial IL [11] optimizes the min-max objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\chi}} \mathbb{E}_{\tau_E}[\log(D_{\boldsymbol{\chi}}(s,a))] + \mathbb{E}_{\tau_{\boldsymbol{\theta}}}[\log(1 - D_{\boldsymbol{\chi}}(s,a))]. \tag{1}$$

Maximizing (1) with respect to $\boldsymbol{\chi}$ is effectively an inverse RL step where a reward function, $r_{\boldsymbol{\chi}}(s,a) = -\log(1 - D_{\boldsymbol{\chi}}(s,a))$, is inferred by leveraging τ_E and $\tau_{\boldsymbol{\theta}}$. Minimizing (1) with respect to $\boldsymbol{\theta}$ is an RL step, where the agent aims to minimize its expected cost. Optimizing (1) is equivalent to minimizing $\mathbb{D}_{\text{JS}}(\rho_{\pi_{\boldsymbol{\theta}}}(s,a) \| \rho_{\pi_E}(s,a))$ [37].

IV. EVENT-BASED PERCEPTION FOR IMITATION WITH VISUAL MISMATCH

Given a target-POMDP and a source-POMDP (i.e. expert), we can categorize the observation space $\mathcal{X}$ into two distinct components: (i) *goal-relevant information*, necessary for successful task completion, and (ii) *visual distractors*, which do not contribute to task execution. Following [7], we define $\mathcal{X}$ as $\mathcal{X} = (\bar{\mathcal{X}}, \hat{\mathcal{X}})$, where $\bar{\mathcal{X}}$ represents goal-relevant information that remains consistent across the source-POMDP and target-POMDP, while $\hat{\mathcal{X}}$ denotes visual distractors that vary across domains. Accordingly, the observations in the source and target domains can be expressed as $x_t^S = (\bar{x}_t, \hat{x}_t^S)$ and $x_t^T = (\bar{x}_t, \hat{x}_t^T)$. The objective of our approach is to remove the distracting information $(\hat{x}_t^T, \hat{x}_t^S \in \hat{\mathcal{X}})$ while preserving the goal-relevant information $(\bar{x}_t \in \bar{\mathcal{X}})$. We summarize our solution in the following paragraphs.

a) Event-based image transformation: Departing from previous work which enforces domain invariance in a learned feature space $\mathcal{Z}$ [3], [5], [7], we propose to extract goal-relevant information $\bar{\mathcal{X}}$ directly from the observation space $\mathcal{X}$. This is accomplished by introducing a transformation $\zeta : \mathcal{X}^2 \to \bar{\mathcal{X}}$, that takes two consecutive RGB frames as input and returns the corresponding event representation as output. Specifically, in $\bar{x}_t$, each pixel independently encodes an event $E_i = (u, v, t_i, p_i)$ (or the lack of it) when the brightness change exceeds a predefined threshold. Here, (u,v) are the pixel indices, t_i is the time step, and $p_i \in \{+1, -1\}$ represents the polarity of intensity change. An event is triggered when:

$$|L_t(u,v) - L_{t-1}(u,v)| \geq C,$$

where $L_t(u,v) = \log(I_t(u,v))$ represents the logarithm of the pixel intensity $I(u,v)$ at time t and C is a predefined threshold. As a result, we obtain $\bar{x}_t = \zeta(x_t, x_{t-1})$ as

$$\bar{x}_t(u,v) = \begin{cases} +1 & \text{if } L_t(u,v) - L_{t-1}(u,v) \geq C, \\ -1 & \text{if } L_t(u,v) - L_{t-1}(u,v) \leq -C, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that in order to avoid fully discarding the whole static content when the camera view is static, we simulate a camera shift by padding x_t by one pixel, i.e., $\bar{x}_t(u,v) = \zeta(x_t(u,v), x_{t-1}(u+1, v+1))$. Importantly, the event transformation in (2) is invariant across multiple manifolds, including brightness and low-frequency image details, which often contain significant visual distractors. In the following, we formally show this property for both *affine transformation* and *low-frequency transformation*.

Consider the brightness variation described by the affine transformation $I' = \alpha I + \beta$. By plugging I' in (2) we obtain

$$L'_t - L'_{t-1} = \log(\alpha I_t + \beta) - \log(\alpha I_{t-1} + \beta) = \log\left(\frac{\alpha I_t + \beta}{\alpha I_{t-1} + \beta}\right).$$

Assuming $\alpha I >> \beta$ yields

$$\approx \log\left(\frac{\alpha I_t}{\alpha I_{t-1}}\right) = \log\left(\frac{I_t}{I_{t-1}}\right) = \log(I_t) - \log(I_{t-1}) = L_t - L_{t-1}.$$

Furthermore, for low frequency changes we obtain

$$\zeta(I_t, I_{t-1} + \eta) = \zeta(I_t, I_{t-1}), \quad \text{if} \quad C - \left|\log \frac{I_t}{I_{t-1}}\right| > \eta,$$

denoting robustness and domain-invariance with respect to η. It is important to note that such properties are not present in any of the popular image transformations (such as greyscale conversion or edge filtering).

Provided the goal-relevant information $\bar{x}_t$, invariant between the source and target POMDPs, the V-IfO problem with mismatches is effectively reduced to a standard V-IfO problem which can be solved with any SOTA algorithm. Due to its improved computational efficiency, we integrate our approach with LAIfO [2] and refer to this event-based version as EB-LAIfO. We introduce the full EB-LAIfO pipeline in the following paragraph.

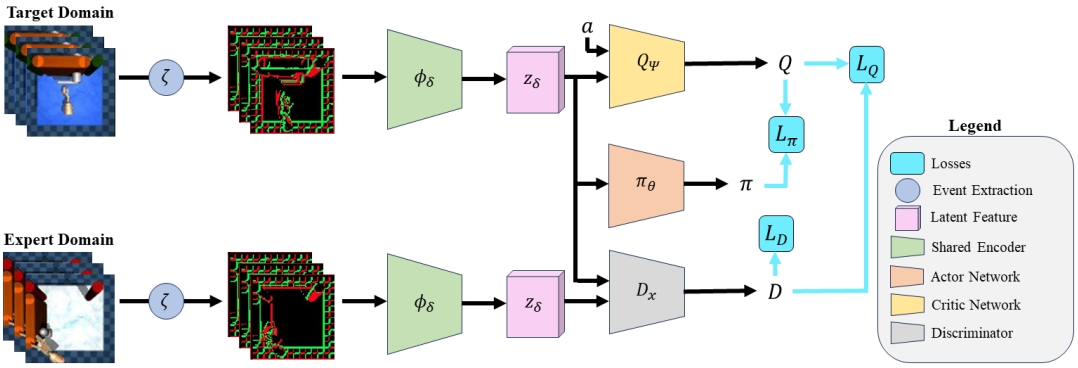
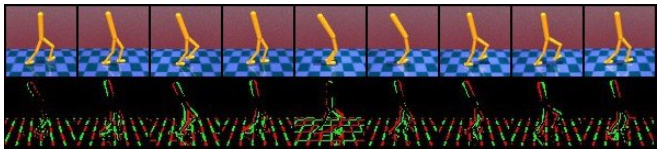
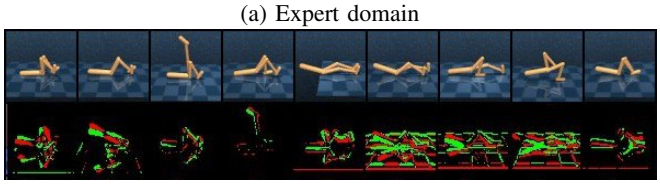


Fig. 1: Summary of EB-LAIfO. Given the RGB sequence, the corresponding events stream is extracted following eq. (2). A feature extractor network ϕ_δ is used to generate the latent features z_δ used by both the Q-function $Q_{\boldsymbol{\psi}}$ and discriminator D_χ for the imitation problem. The discriminator D_χ is trained as in (4) and returns the reward function $r_{\boldsymbol{\chi}}$ which is then maximized through an RL step. The RL step is described in (3) and follows the Deep Deterministic Policy Gradient (DDPG) pipeline [38]. Our feature extractor network ϕ_δ is trained jointly with the Q-function and extracts goal-relevant information directly from the events stream.



(a) Expert domain



(b) Target Domain

Fig. 2: Examples scenes from the expert and target domain, both in original RGB space and obtained event spaced (green is positive event, red is negative).

b) Event-based latent adversarial imitation from observations: Suppose we have access to a video of an expert demonstrating a task, denoted as $\tau_E = x^S_{0:T}$. Our goal is to perform AIL, as introduced in Section III-.0.c, directly from this video. To address this challenge, we first estimate a latent variable $z \in \mathcal{Z}$ from the observation space and then execute the entire AIL pipeline within this latent space. However, naively applying this approach in our setting results in poor performance due to the inherent visual mismatch between the source and target domains, coupled with the absence of expert action data in the dataset. These challenges make latent variable estimation significantly more difficult [2], [7] (see also Section 6.10 in [39]). To mitigate these issues, we first apply our event-based transformation, introduced in (2), to process all observation streams collected from both the source and target POMDPs. Then, within our event-based observation space $\bar{\mathcal{X}}_{\text{EB}}$, we define a feature extractor $\phi_{\boldsymbol{\delta}} : \bar{\mathcal{X}}^d_{\text{EB}} \to \mathcal{Z}$, which takes as input a stack of $d \in \mathbb{N}$ observations such that $z = \phi_{\boldsymbol{\delta}}(\bar{x}_{t^-:t})$, where $t - t^- + 1 = d$. The latent variable $z \in \mathcal{Z}$ is estimated directly in $\bar{\mathcal{X}}_{\text{EB}}$ by training the feature extractor $\phi_{\boldsymbol{\delta}}$ jointly with the critic networks $Q_{\boldsymbol{\psi}_k}$ $(k = 1, 2)$. Specifically, we minimize the following loss

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\delta},\boldsymbol{\psi}_k}(\mathcal{B}) &= \mathbb{E}_{(z,a,z')\sim\mathcal{B}}[(Q_{\boldsymbol{\psi}_k}(z,a) - y)^2], \\ y &= r_{\boldsymbol{\chi}}(z,z') + \gamma \min_{k=1,2} Q_{\bar{\boldsymbol{\psi}}_k}(z',a'), \end{aligned} \tag{3}$$

where a is an action stored in the agent buffer $\mathcal{B}$ and which is used by the agent to interact with the environment, while $a' = \pi_{\boldsymbol{\theta}}(z') + \epsilon$ where $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma^2), -c, c)$ is a clipped exploration noise with c the clipping parameter and $\mathcal{N}(0, \sigma^2)$ a univariate normal distribution with zero mean and σ standard deviation. Finally, $\bar{\boldsymbol{\psi}}_1$ and $\bar{\boldsymbol{\psi}}_2$ are the slow-moving weights for the target Q networks.

The reward function $r_{\boldsymbol{\chi}}(z,z')$ in (3) is defined as $r_{\boldsymbol{\chi}}(z,z') = -\log\big(1 - D_{\boldsymbol{\chi}}(z,z')\big)$, where $D_{\boldsymbol{\chi}} : \mathcal{Z} \times \mathcal{Z} \to [0,1]$ is a discriminator function trained to optimize the following loss, as in (1):

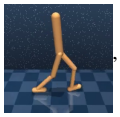
$$\begin{aligned} \max_{\boldsymbol{\chi}} \; & \mathbb{E}_{(z,z')\sim\mathcal{B}_E(\tau_E)}[\log(D_{\boldsymbol{\chi}}(z,z'))] \\ & + \mathbb{E}_{(z,z')\sim\mathcal{B}(\tau_{\boldsymbol{\theta}})}[\log(1 - D_{\boldsymbol{\chi}}(z,z'))]. \end{aligned} \tag{4}$$

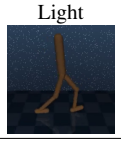
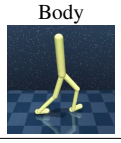
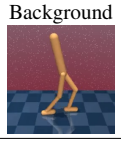
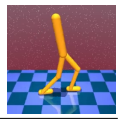
Here, $\mathcal{B}_E$ and $\mathcal{B}$ are the replay buffers storing τ_E and $\tau_{\boldsymbol{\theta}}$, corresponding to the expert and learning agent observations, respectively. In this final step, both expert and learning agent observations are first preprocessed using our event-based transformation such that $\bar{x}_t = \zeta(x^S_t, x^S_{t-1})$ and $\bar{x}_t = \zeta(x^T_t, x^T_{t-1})$ and mapped onto $\mathcal{X}_{\text{EB}}$. Furthermore, they are embedded as $z = \phi_{\boldsymbol{\delta}}(\bar{x}_{t^-:t})$ to formulate the problem compactly in $\mathcal{Z}$. To streamline notation, we write $(z,z') \sim \mathcal{B}(\tau_{\boldsymbol{\theta}})$ and $(z,z') \sim \mathcal{B}_E(\tau_E)$. We provide a schematic visualization of the full pipeline in Fig. 1.

V. EXPERIMENTS

In this section, we first demonstrate how EB-LAIfO effectively handles various types of visual mismatches in the V-IfO setting (Sec. V-A); and then showcase how EB-LAIfO is more effective than other algorithms at facilitating learning in challenging robotic manipulation tasks with sparse rewards

TABLE I: Summary of the experiments on the DeepMind control suite [9]. We train all the algorithms for 500,000 steps. The learned policies are evaluated based on average return over 10 episodes. We report the mean and standard deviation of the final return across 5 seeds and **highlight** the best performance. These results highlight the effectiveness of EB-LAIfO in handling light and color mismatches compared to the tested baselines.

Target Env:  , Expert Performance = 950

Source Env	Light 	Body 	Floor 	Background 	Full 
EB-LAIfO (ours)	**677 ± 285**	**854 ± 125**	**603 ± 271**	**858 ± 108**	**898 ± 48**
C-LAIfO [7]	202 ± 287	778 ± 155	123 ± 55	441 ± 328	203 ± 78
LAIfO [2] w/ data aug	63 ± 28	741 ± 102	353 ± 200	69 ± 48	142 ± 230
Canny-Edge-LAIfO	52 ± 24	304 ± 179	28 ± 5	703 ± 126	129 ± 19
Grayscale-LAIfO	24 ± 3	102 ± 11	28 ± 4	26 ± 1	23 ± 6
DisentanGAIL [5]	26 ± 4	425 ± 247	46 ± 16	25 ± 5	26 ± 6
PatchAIL [1] w/ data aug	22 ± 8	170 ± 65	14 ± 1	30 ± 6	40 ± 19

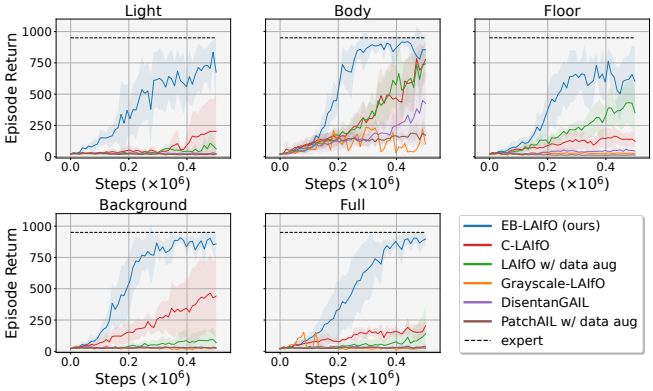


Fig. 3: Learning curves for the results in Table I. Plots show the average return per episode and the standard deviation across seeds as a function of training steps.

and realistic visual inputs (Sec. V-B). Finally, in Sec. V-C, we study the robustness of our approach to noise in the event stream. Unless specified, we use DDPG [40] to train experts in a fully observable setting and collect 100 episodes of expert videos for imitation.

A. *Visual Imitation from Observations with mismatch*

In this paragraph, we evaluate EB-LAIfO in the V-IfO setting across various visual mismatches and compare it with four baselines: C-LAIfO [7], LAIfO [2], PatchAIL [1], and DisentanGAIL [5]. For the Light experiment, we provide our baselines with brightness transformation as a data augmentation technique, while for the other experiments, we define data augmentation as a color transformation. In addition, we also include a LAIfO baseline using two other transformations, namely RGB-to-grayscale (i.e. 'Grayscale-LAIfO') and canny-edge detection (i.e. 'Canny-Edge-LAIfO'), serving as an alternative to the event-based transformation. The results, summarized in Table I, are further illustrated through learning curves in Fig. 3. In our assessment of performance, we consider both final results and learning efficiency.

The results in Table I demonstrate that EB-LAIfO consistently outperforms all other baselines in handling visual mismatches. Notably, EB-LAIfO reaches near-expert performance in the Full setting (898 ± 48), achieving a fourfold improvement over C-LAIfO (203 ± 78). Furthermore, when the discrepancy in final performance is smaller, such as in the Body setting, we observe that EB-LAIfO converges significantly more efficiently than all other tested methods.

The key reasons for this superior performance are twofold. First, event-based perception retains goal-relevant information more effectively than data augmentation-based techniques, which heavily depend on the types of augmentations and are more susceptible to stochasticity, as augmentations are randomized. Second, since our event-based transformation is a purely mathematical operation that does not require active learning, our approach is inherently more efficient than learning-based methods. For example, the contrastive learning step in C-LAIfO requires extensive training and data augmentation to extract domain-invariant features, whereas event-based transformations enhance this invariance by filtering out visual distractors directly from the pixel space.

This advantage leads to improved performance and sample efficiency, particularly in environments with strong visual mismatches, such as Background (858 ± 108 for EB-LAIfO vs. 441 ± 328 for C-LAIfO) and Floor (603 ± 271 for EB-LAIfO vs. 123 ± 55 for C-LAIfO). Furthermore, the limitations of traditional methods are evident, as DisentanGAIL, LAIfO and PatchAIL fail to handle domain shifts effectively, achieving low scores for most of the mismatches.

Notably, other transformation techniques fail to produce robust representations. In fact, approaches such as RGB-to-grayscale or Canny edge detection operate on the image brigthness and therefore are not invariant to changes in the

TABLE II: Summary of the experiments on the Adroit platform for dynamic dexterous manipulation [10]. The Door-Light and Door-Color experiments consider (5a) as the source-POMDP and (5b) and (5c) as the respective target-POMDPs. Similarly, the Hammer-Light and Hammer-Color experiments consider (5d) as source-POMDP and (5e) and (5f) as the respective target-POMDPs. We use the VRL3 in [41] to train expert policies and collect 100 episodes of expert data. All the algorithms are trained for 10^6 steps. The learned policies are evaluated based on average return over 10 episodes. We report the mean and standard deviation of the final return across 4 seeds and **highlight** the best performance.

	Door		Hammer	
	Light	Color	Light	Color
Expert	170		184	
RL+EB-LAIfO (ours)	111 ± 59	$\mathbf{106 \pm 48}$	$\mathbf{174 \pm 10}$	$\mathbf{177 \pm 9}$
RL+C-LAIfO [7]	$\mathbf{150 \pm 5}$	79 ± 81	99 ± 76	118 ± 74
RL+LAIfO [2]	35 ± 64	68 ± 71	82 ± 85	-2 ± 0.0

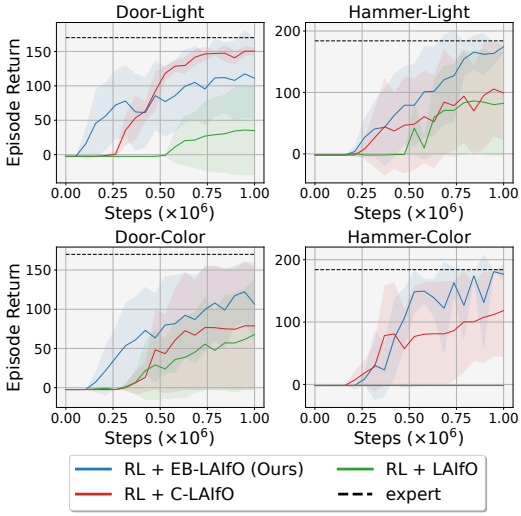
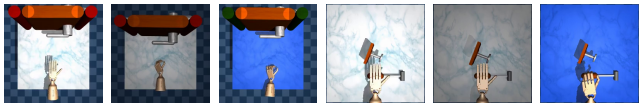
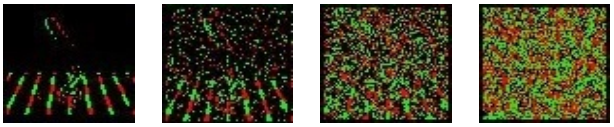


Fig. 4: Learning curves for the results described in Table II. Plots show the average return per episode and the standard deviation across seeds as a function of training steps.



(a) (b) (c) (d) (e) (f)

Fig. 5: Adroit environments used for the experiments in Table II.



(a) $\sigma = 0.01$ (b) $\sigma = 0.1$ (c) $\sigma = 0.2$ (d) $\sigma = 0.5$

Fig. 6: Visualization of the resulting event stream with varying noise standard deviation level σ.

scene appearance. Moreover, as they only perform static transformations of the input image re-encoding appearance information, these methods do not address the underlying source of visual mismatches and cannot disentangle motion dynamics from visual style. In contrast, our event-based transformation fundamentally alters the sensory representation by encoding temporal intensity gradients. This enables the agent to focus on task-relevant transients while discarding static appearance features that are most affected by domain shifts.

These results confirm that EB-LAIfO's event-based perception framework is significantly more robust and sample-efficient than other learning-based approaches, making it a promising solution for real-world IL in visually mismatched environments.

B. Dexterous manipulation experiments

In this section, we evaluate our algorithm on a series of challenging robotic manipulation tasks from the Adroit platform for dynamic dexterous manipulation [10]. These experiments demonstrate how the reward $r_{\boldsymbol{\chi}}$, learned by EB-LAIfO from expert videos, can be effectively combined with a sparse reward $\mathcal{R}$, collected by the agent through interaction with the environment, to enhance learning efficiency. The RL problem aims to maximize the total reward

$$\mathcal{R}_{\text{tot}} = \mathcal{R}(s_t, a_t) + r_{\boldsymbol{\chi}}(z_t, z_{t+1}), \tag{5}$$

where $r_{\boldsymbol{\chi}}(z, z') = -\log\big(1 - D_{\boldsymbol{\chi}}(z, z')\big)$ with $D_{\boldsymbol{\chi}}(z, z')$ trained to optimize (4). This approach is particularly relevant for robotic tasks, where sparse rewards are often the most feasible option in real-world settings. However, relying solely on sparse rewards can make learning challenging and inefficient. In this context, leveraging expert videos can significantly enhance efficiency.

To evaluate the effectiveness of EB-LAIfO, we compare it against C-LAIfO and the standard LAIfO algorithm on the Adroit platform for dynamic dexterous manipulation [10]. All these methods employ an encoder to process pixel observations, extracting embeddings in $\mathcal{Z}$, which are then concatenated with robot sensory observations. Notably, the expert's sensory observations are not used in the imitation process, as we assume access only to expert videos. Consequently, in these experiments, we seek to maximize $\mathcal{R}_{\text{tot}}$ in (5), rather than just $r_{\boldsymbol{\chi}}$, as in standard imitation learning.

The results, summarized in Table II, highlight the advantages of EB-LAIfO in handling visual mismatches. Across all conditions, EB-LAIfO consistently outperforms LAIfO and exhibits strong performance compared to C-LAIfO. In the Hammer tasks, EB-LAIfO achieves near-expert performance with 174 ± 10 in the Light setting and 177 ± 9 in the Color setting, demonstrating its robustness in retaining task-relevant features despite visual perturbations. In contrast, C-LAIfO shows reduced performance in these settings (99 ± 76 and 118 ± 74, respectively), indicating that it struggles to fully bridge domain gaps introduced by color variations.

In the Door tasks, C-LAIfO achieves the best performance

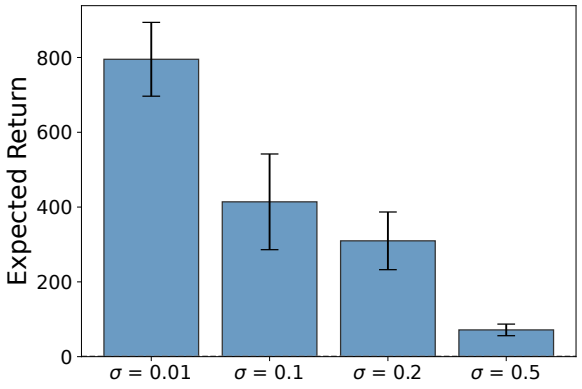


Fig. 7: Ablation study with different σ noise values. Each policy is evaluated as in Table I. We report mean and standard deviation across the last 100 evaluations for each run.

in the Light setting (150 ± 5), slightly surpassing EB-LAIfO (111 ± 59), suggesting that C-LAIfO may be more effective in handling brightness-based domain shifts in this scenario. However, EB-LAIfO significantly outperforms C-LAIfO in the more challenging Door-Color setting (106 ± 48 vs. 79 ± 81), reinforcing the idea that event-based transformations are particularly effective at mitigating challenging color-based domain mismatches. LAIfO exhibits the weakest performance across all experiments, failing to generalize effectively under domain shifts. In the Hammer-Color setting, it even fails to learn a meaningful policy, achieving a negative return (-2 ± 0.0), while in the Door-Light setting, it only reaches 35 ± 64, far below the other methods.

These results confirm that EB-LAIfO provides a robust and sample-efficient approach for imitation in visually mismatched environments, particularly when handling complex domain shifts such as those introduced by color variations.

C. Event noise ablation

To assess the robustness of our approach, as well as proving its applicability for real event cameras, we perform an ablation study injecting Gaussian noise to simulate real-world sensor imperfections and environmental interference. In these experiments, our transformation function ζ becomes

$$\bar{x}_t(u,v) = \begin{cases} +1 & \text{if } L_t(u,v) - L_{t-1}(u,v) + \mathcal{N}(0,\sigma^2) \geq C, \\ -1 & \text{if } L_t(u,v) - L_{t-1}(u,v) + \mathcal{N}(0,\sigma^2) \leq -C, \\ 0 & \text{otherwise,} \end{cases}$$

where σ represents the (fixed) standard deviation of the additive Gaussian noise. While there are other types of noise occurring in real-world scenarios, Gaussian noise provides a reasonable approximation of the combined effect of multiple noise sources, making it an appropriate choice for our robustness evaluation.

We consider four noise levels with $\sigma_{1,..,4}$ set respectively to $0.01, 0.1, 0.2$ and 0.5. Resulting event samples are visualized in Fig. 6. The final results, illustrated in Fig 7, reveal that our approach maintains robust performance up to moderate noise levels ($\sigma = 0.2$), with outcomes comparable to the noise-free baseline. While performance degradation becomes evident at higher noise intensities, it's worth noting that such extreme noise levels (as illustrated in Fig. 6) severely corrupt the sensor information to an extent that would be unrealistic for properly functioning event cameras in practical applications.

VI. Conclusion

In this paper, we introduce a novel approach to V-IfO that addresses the fundamental challenge of visual domain gaps between demonstration and deployment environments. By drawing inspiration from biological vision systems and event camera technology, we developed a lightweight, event-inspired perception module that inherently filters out appearance-based distractors while preserving the motion dynamics critical for successful imitation. Our experiments demonstrate that discretized temporal gradients provide a robust alternative to conventional learning-based methods. This approach eliminates the need for expensive domain randomization or handcrafted data augmentation strategies that have limited current SOTA methods.

A limitation of our approach arises in scenarios where appearance features, rather than just motion dynamics, provide critical contextual cues for task completion. Another limitation of our current implementation lies in the fidelity of our synthetic event stream generation from RGB videos. Our model does not fully capture all secondary effects of real-world event camera systems, such as complex noise models and asynchronous data streams. While sufficient as a proof-of-concept, real event cameras capture temporal information at microsecond resolution, which our conversion process cannot fully replicate. An interesting direction for future work is to apply our method in real-world applications, integrating either standard RGB or event cameras. Additionally, we believe our approach offers a promising alternative for bridging another common domain gap, namely the one between synthetic and real-world visual data. From an algorithmic perspective, exploring different representation learning techniques for extracting goal-relevant features from event-based images represents another exciting avenue. Methods that better exploit the unique characteristics of event-based data streams could further enhance performance and improve efficiency, not only for IL but also for RL in POMDP settings.

ACKNOWLEDGMENT

This work was supported by the German Federal Ministry for Economic Affairs and Energy (BMWE) within the project "NXT GEN AI METHODS".

REFERENCES

[1] M. Liu, T. He, W. Zhang, Y. Shuicheng, and Z. Xu, "Visual imitation learning with patch rewards," in *International Conference on Learning Representations*, 2022.

[2] V. Giammarino, J. Queeney, and I. Paschalidis, "Adversarial imitation learning from visual observations using latent information," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=ydPHjgf6h0

[3] B. C. Stadie, P. Abbeel, and I. Sutskever, "Third-person imitation learning," *arXiv preprint arXiv:1703.01703*, 2017.

[4] R. Okumura, M. Okada, and T. Taniguchi, "Domain-adversarial and-conditional state space model for imitation learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5179–5186.

[5] E. Cetin and O. Celiktutan, "Domain-robust visual imitation learning with mutual information constraints," in *International Conference on Learning Representations*, 2020.

[6] S. Choi, S. Han, W. Kim, J. Chae, W. Jung, and Y. Sung, "Domain adaptive imitation learning with visual observation," in *Advances in Neural Information Processing Systems*, 2023.

[7] V. Giammarino, J. Queeney, and I. C. Paschalidis, "Visually robust adversarial imitation learning from videos with contrastive learning," *arXiv preprint arXiv:2407.12792*, 2024.

[8] H. Kolb, E. Fernandez, and R. Nelson, "The organization of the retina and visual system," *Salt Lake City (UT): University of Utah Health Sciences Center*, 1995.

[9] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq *et al.*, "Deepmind control suite," *arXiv preprint arXiv:1801.00690*, 2018.

[10] V. Kumar, "Manipulators and manipulation in high dimensional spaces," Ph.D. dissertation, University of Washington, Seattle, 2016. [Online]. Available: https://digital.lib.washington.edu/researchworks/handle/1773/38104

[11] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[13] S. Russell, "Learning agents for uncertain environments," in *Proceedings of the Annual Conference on Computational Learning Theory*, 1998, pp. 101–103.

[14] A. Y. Ng, S. Russell *et al.*, "Algorithms for inverse reinforcement learning." in *International Conference on Machine Learning*, vol. 1, 2000, p. 2.

[15] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *International Conference on Machine Learning*, 2004, p. 1.

[16] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *AAAI Conference on Artificial Intelligence*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[17] T. Gangwani, J. Lehman, Q. Liu, and J. Peng, "Learning belief representations for imitation learning in POMDPs," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 1061–1071.

[18] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn, "Visual adversarial imitation learning using variational models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3016–3028, 2021.

[19] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," *arXiv preprint arXiv:1807.06158*, 2018.

[20] C. Yang, X. Ma, W. Huang, F. Sun, H. Liu, J. Huang, and C. Gan, "Imitation learning from observations by minimizing inverse dynamics disagreement," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] Z. Cheng, L. Liu, A. Liu, H. Sun, M. Fang, and D. Tao, "On the guaranteed almost equivalence between imitation learning from observation and demonstration," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[24] K. Kim, Y. Gu, J. Song, S. Zhao, and S. Ermon, "Domain adaptive imitation learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5286–5295.

[25] D. S. Raychaudhuri, S. Paul, J. Vanbaar, and A. K. Roy-Chowdhury, "Cross-domain imitation from observations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8902–8912.

[26] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1118–1125.

[27] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[28] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine, "Avid: Learning multi-stage tasks via pixel-level translation of human videos," *Robotics: Science and Systems XVI*, 2020.

[29] V. Giammarino, J. Queeney, L. C. Carstensen, M. E. Hasselmo, and I. C. Paschalidis, "Opportunities and challenges from using animal videos in reinforcement learning for navigation," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 9056–9061, 2023.

[30] J. Z. Zhang, S. Yang, G. Yang, A. L. Bishop, S. Gurumurthy, D. Ramanan, and Z. Manchester, "Slomo: A general system for legged robot motion imitation from casual videos," *IEEE Robotics and Automation Letters*, 2023.

[31] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[32] D. A. Silva, K. Smagulova, A. Elsheikh, M. E. Fouda, and A. M. Eltawil, "A recurrent yolov8-based framework for event-based object detection," *Frontiers in Neuroscience*, vol. 18, p. 1477979, 2025.

[33] B. Forrai, T. Miki, D. Gehrig, M. Hutter, and D. Scaramuzza, "Event-based agile object catching with a quadrupedal robot," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 177–12 183.

[34] A. Bhattacharya, M. Cannici, N. Rao, Y. Tao, V. Kumar, N. Matni, and D. Scaramuzza, "Monocular event-based vision for obstacle avoidance with a quadrotor," *arXiv preprint arXiv:2411.03303*, 2024.

[35] T. Kim, H. Cho, and K.-J. Yoon, "Frequency-aware event-based video deblurring for real-world motion blur," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 966–24 976.

[36] R. Pellerito, M. Cannici, D. Gehrig, J. Belhadj, O. Dubois-Matra, M. Casasco, and D. Scaramuzza, "Deep visual odometry with events and frames," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8966–8973.

[37] S. K. S. Ghasemipour, R. Zemel, and S. Gu, "A divergence minimization perspective on imitation learning methods," in *Proceedings of the Conference on Robot Learning*. PMLR, 2020, pp. 1259–1277.

[38] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*. Pmlr, 2014, pp. 387–395.

[39] V. Giammarino, "On the use of expert data to imitate behavior and accelerate reinforcement learning," Ph.D. dissertation, Boston University, 2024.

[40] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[41] C. Wang, X. Luo, K. Ross, and D. Li, "Vrl3: A data-driven framework for visual deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 974–32 988, 2022.