ACTIVE ACQUISITION FOR MULTIMODAL TEMPORAL DATA: A CHALLENGING DECISION-MAKING TASK

Jannik Kossen[1*†] **Cătălina Cangea**[2†] **Eszter Vértes**[2] **Andrew Jaegle**[2]
Viorica Patraucean[2] **Ira Ktena**[2] **Nenad Tomasev**[2] **Danielle Belgrave**[2]

[1] OATML, Department of Computer Science, University of Oxford
[2] DeepMind

ABSTRACT

We introduce a challenging decision-making task that we call *active acquisition for multimodal temporal data* (A2MT). In many real-world scenarios, input features are not readily available at test time and must instead be acquired at significant cost. With A2MT, we aim to learn agents that actively select which modalities of an input to acquire, trading off acquisition cost and predictive performance. A2MT extends a previous task called *active feature acquisition* to temporal decision making about high-dimensional inputs. Further, we propose a method based on the Perceiver IO architecture to address A2MT in practice. Our agents are able to solve a novel synthetic scenario requiring practically relevant cross-modal reasoning skills. On two large-scale, real-world datasets, Kinetics-700 and AudioSet, our agents successfully learn cost-reactive acquisition behavior. However, an ablation reveals they are unable to learn to learn adaptive acquisition strategies, emphasizing the difficulty of the task even for state-of-the-art models. Applications of A2MT may be impactful in domains like medicine, robotics, or finance, where modalities differ in acquisition cost and informativeness.

1 INTRODUCTION

In making a clinical diagnosis, the medical professional must carefully choose a specific set of tests to diagnose the patient quickly and correctly. It is of crucial importance to choose the right test at the right time, and tests should only be performed when useful, as they may otherwise cause unnecessary patient discomfort or financial expense. Recently, large-scale datasets of medical treatment records have become available (Hyland et al., 2020; Johnson et al., 2020). They may potentially facilitate improvements in medical domain knowledge and patient care, for example by allowing us to learn which tests to perform when. While prior work has demonstrated that machine learning can be used to inform complex diagnoses from simple measurements, e.g. (Liotta et al., 2003; Diaz-Pinto et al., 2022), treatment records present a significant modelling challenge as they contain temporally sparse observations from high-dimensional modalities, e.g. X-Rays, MRIs, blood tests, or genetic data.

Prior work in *active feature acquisition* (AFA) (Greiner et al., 2002; Melville et al., 2004; Ling et al., 2004; Zubek et al., 2004; Sheng & Ling, 2006; Saar-Tsechansky et al., 2009) has similarly considered the cost of feature acquisition at test time on a per-datum basis: given a test input for which all features are missing initially, which features should one acquire to best trade off predictive performance and the cost of feature acquisition? Concurrent methods in AFA often use a Bayesian experimental design approach (Lindley, 1956; Chaloner & Verdinelli, 1995; Sebastiani & Wynn, 2000), acquiring features that maximise the expected information gain with respect to the prediction (Ma et al., 2018; Li & Oliva, 2021; Lewis et al., 2021). Alternatively, AFA can be phrased as a reinforcement learning task where agents optimize the trade-off objective directly (Shim et al., 2018; Kachuee et al., 2019; Zannone et al., 2019; Janisch et al., 2019; 2020; Yin et al., 2020). Notably, prior work in AFA assumes static data: although acquisitions are sequential, feature values do not evolve along a temporal dimension. Furthermore, the features themselves usually correspond to low-dimensional observations, i.e. single values in a tabular dataset.

In this work, we propose *active acquisition for multimodal temporal data* (A2MT). Taking the above medical setting as motivation, we extend the familiar setting of AFA in two key ways: (1) We assume that inputs are sequences that evolve temporally. Our agents will need to learn not only

[*]Work done while interning at DeepMind.
[†]Correspondence to jannikkossen@cs.ox.ac.uk and ccangea@deepmind.com.

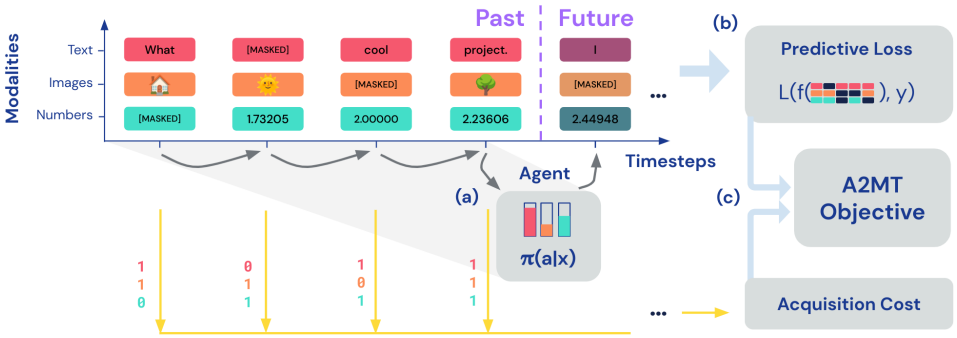

Figure 1: In many practical applications, features are not available a priori at test time and have to be acquired at a real-world cost to allow for the prediction of an associated label. In *Active Acquisition for Multimodal Temporal Data*, we aim to learn agents that efficiently acquire for multimodal temporal inputs: (a) at each timestep, the agent decides which modalities of the input it acquires, paying a per-modality acquisition cost; (b) then, a separate model predicts given the sparse sequence of observations; (c) lastly, the agent gets rewarded for low prediction loss and small acquisition cost.

which features to acquire, but also *when* to acquire them. (2) We no longer assume that inputs are unimodal and low-dimensional. Instead, we assume that each input comprises a collection of high-dimensional modalities, and that acquisitions are made for entire modalities at each timestep. With these extensions, A2MT generalizes AFA and reduces the gap to practical applications that are often both temporal and multimodal. A2MT can also find application outside the medical domain (cf. §6).

To study A2MT in a controlled environment, we propose a set of synthetic scenarios of increasing difficulty that are temporal *and* multimodal, both key requirements for A2MT. Further, we propose to study A2MT on audio-visual datasets, concretely AudioSet (Gemmeke et al., 2017) and Kinetics-700 2020 (Smaira et al., 2020). These provide a challenging testbed for A2MT and avoid some of the complications of working with medical data. We propose a method based on Perceiver IO (Jaegle et al., 2021)—a modality-agnostic architecture that can be applied directly to a large variety of real-world inputs—and we explore different reinforcement learning techniques to train the agent. Our method is able to solve a subset of the synthetic tasks we propose and provides reasonable performance on the real-world datasets. However, further investigation reveals that our method can ultimately be outperformed by a non-adaptive ablation.

In summary, our contributions are:

- We introduce the *active acquisition for multimodal temporal data* (A2MT) scenario (§2).
- We suggest both synthetic and real-world datasets that motivate A2MT and allow for convenient benchmarking of methods (§3).
- We propose a novel method based on Perceiver IO to tackle A2MT (§4), provide a thorough empirical study (§5), and discuss key areas of improvement for future work (§7).

2 Active Acquisition for Multimodal Temporal Data

With A2MT (Fig. 1), we study strategies for cost-sensitive acquisition when data is both multimodal and temporal. We train an agent with policy π that makes a binary acquisition decision per modality and timestep. After a fixed number of timesteps, a model f makes a prediction given the sparse sequence of acquired observations. The agent then observes a reward that consists of two terms: (1) The agent gets rewarded in proportion to the negative predictive loss given the observations. (2) Acquisitions come at a (modality-specific) cost and the agent is penalized for each acquisition made. Term (1) compels the agent to acquire often, as additional observations should improve the quality of the predictions of f. However, this increases the penalty incurred from term (2). The agent therefore needs to trade off acquisition cost and prediction reward, learning which modalities in the input are worth acquiring and when it is worth acquiring them.

To make the most of a limited acquisition budget, we aim to learn agents with individualized acquisition strategies that adapt to the sequences at hand. For example, agents may learn to make use of interactions between modalities: a past observation in one modality (e.g. a suspicious value in a blood test) may lead to the acquisition of another modality (e.g. a more specialised test), achieving both accuracy and cost-efficiency. A2MT requires models that can successfully accommodate sparse, multimodal, and temporal data. This is challenging for state-of-the-art models, even without the additional complexities of active selection required for successful A2MT.

In A2MT (Alg. 1), we assume access to a fully observed training set of input-output pairs $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$. For classification tasks, for example, y_i are labels, $y_i \in (1, \ldots, C)$. Each input $\boldsymbol{x}_i$ is a sequence of observations $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T})$. At each timestep t, the observation $\boldsymbol{x}_{i,t}$ decomposes into M modalities $\boldsymbol{x}_{i,t} = (\boldsymbol{x}_{i,t,1}, \ldots, \boldsymbol{x}_{i,t,M})$; each modality may be high-dimensional, $\boldsymbol{x}_{i,t,m} \in \mathbb{R}^{d_m}$. For example, $\boldsymbol{x}_{i,t,m}$ could be a single frame in a video where we collapse the image height h, width w, and color channels c into a single axis with dimensionality $d_m = h \cdot w \cdot c$. We focus on a single sample and drop the leading axis, $\boldsymbol{x} = \boldsymbol{x}_i$, to avoid notational clutter. We use colons to indicate 'slices' of inputs along a particular axis, e.g. $\boldsymbol{x}_{1:t-1} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1})$.

At each timestep $t \in (1, \ldots, T)$, we obtain a set of acquisition decisions across modalities, or actions $\boldsymbol{a}_t = (a_{t,1}, \ldots, a_{t,M})$, by sampling from the agent policy $\boldsymbol{a}_t \sim \pi(\cdot|\tilde{\boldsymbol{x}}_{1:t-1}, a_{1:t-1}; \boldsymbol{\theta})$. Here, $a_{t,m} \in (0, 1)$ is a binary indicator of whether modality m was acquired at time t. We write $\tilde{\boldsymbol{x}}$ instead of $\boldsymbol{x}$ to highlight the fact that the inputs may contain missing entries, and $\boldsymbol{\theta}$ are the trainable agent parameters. Here, π gives the joint probability over all possible acquisition decisions at that timestep, including extremes such as acquiring all or none of the modalities, i.e. $\pi(\boldsymbol{a}_t|\tilde{\boldsymbol{x}}_{1:t-1}, a_{1:t-1}; \boldsymbol{\theta}) \in [0, 1]^M$. At each timestep t and for each modality m, we acquire $\boldsymbol{x}_{m,t}$ only if $a_{m,t} = 1$. We summarize the set of all actions across timesteps as $\boldsymbol{a} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T)$.

Algorithm 1 A2MT

Inputs: Test input $\boldsymbol{x}$, agent π, model f.
for $t = 1$ to T **do**
  Sample $a_t \sim \pi(\cdot|\tilde{\boldsymbol{x}}_{1:t-1}, \boldsymbol{a}_{1:t-1}; \boldsymbol{\theta})$.
  for $m = 1$ to M **do**
    if $a_{t,m} = 1$ **then**
      Acquire: $\tilde{\boldsymbol{x}}_{t,m} \leftarrow \boldsymbol{x}_{t,m}$.
    else
      Do not acquire: $\tilde{\boldsymbol{x}}_{t,m} \leftarrow \emptyset$.
    end if
  end for
end for
Return prediction $f(\tilde{\boldsymbol{x}}_{1:T})$.

Lastly, we assume access to a model f which makes a prediction $\hat{y}$ given an input $\tilde{\boldsymbol{x}}$. After completing the acquisition process for a given test sample, we use this model to predict $f(\tilde{\boldsymbol{x}}_{1:T}) = \hat{y}$. Consequently, we require that f can predict for multimodal inputs with features missing arbitrarily.

We train agents to maximize the following reward:

$$R = \mathbb{E}\left[-C(\boldsymbol{a}) - \mathcal{L}(f(\tilde{\boldsymbol{x}}_{1:T}), y)\right], \quad \text{where} \quad C(\boldsymbol{a}) = \sum_{t=1}^{T} \sum_{m=1}^{M} c_m a_{t,m}. \tag{1}$$

Here, the expectation is over the data $(\boldsymbol{x}, y)$ and actions $\boldsymbol{a}$ from the policy; $C(\boldsymbol{a})$ gives the total cost of acquisition along the sequence; c_m is a modality-specific cost. Equation (1) summarizes the problem of A2MT: trading off acquisition costs against predictive performance for multimodal and temporal inputs. In §4, we detail how we optimize the objective using reinforcement learning.

3 Datasets for A2MT

While the medical domain acts as important practical *motivation* for A2MT, medical datasets such as Hyland et al. (2020); Johnson et al. (2020) usually require domain expertise. While the eventual application of A2MT methods in this domain is crucial, we believe that a mix of synthetic and real *non-medical* data can serve as a widely accessible testbed for the machine learning community to develop robust methodology. We therefore introduce datasets suitable for developing A2MT methods.

3.1 Synthetic Datasets

We introduce three scenarios with varying levels of difficulty and a clear optimal acquisition strategy, designed to test the cross-modal reasoning capabilities of the agents.

Concretely, we create a dataset (cf. Algorithm 2) of fixed-length sequences with two modalities: `counter` and `digit`. The `digit` modality is a sequence of random numbers, e.g. `0131043443`. The `counter` modality repeatedly counts down from a randomly drawn starting value,

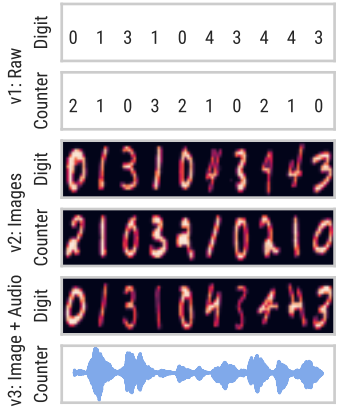

Figure 2: The synthetic scenarios allow for sparse acquisition while keeping perfect accuracy. This requires agents that capable of practically relevant cross-modal reasoning. (Label is 9 in the above.)

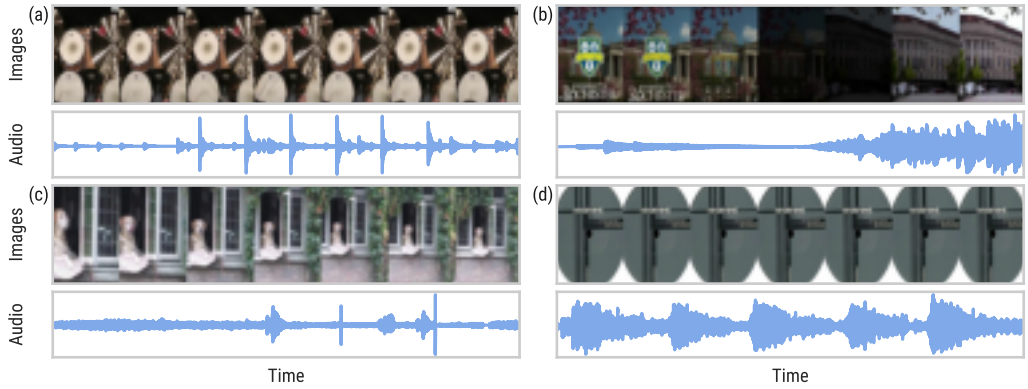

Figure 3: Four example sequences from the AudioSet training set. The labels associated with these inputs are (a) Music, (b) Music, (c) Speech, and (d) Electronic music. The audio signal is often more informative than the images for AudioSet. Inputs are downsampled in the above visualization.

e.g. `2103210210`. The label attached to each sequence is the sum of the `digit` modality at all timesteps where the `counter` modality is zero, e.g. `3+3+3=9` in this case.

To increase complexity of this task, we replace the sequence of raw numbers with a sequence of images, for example MNIST digits (LeCun, 1998), where each raw number is replaced by a matching image from the MNIST dataset. We can also replace the numbers of the `counter` modality with matching *audio* sequences from the spoken digit dataset (Jackson, 2017). We can increase task complexity by adjusting the sequence length and the number of unique symbols for the modalities. Figure 2 visualizes the different synthetic dataset variants.

To solve the synthetic task, agents need to reason about interactions between modalities, a key feature of A2MT. Each modality in the synthetic scenario offers an ideal strategy to save acquisition cost without sacrificing predictive accuracy. (1) The agent can learn to acquire the `digit` modality only if the `counter` modality is zero, because only then is a `digit` relevant for the label. (2) The agent can further learn to *skip* observations in the `counter` modality.

3.2 Audio-Visual Datasets

We propose to apply large-scale audio-visual classification datasets, such as the AudioSet (Gemmeke et al., 2017) or Kinetics-700-2020 (Smaira et al., 2020), to the A2MT setting. These datasets are multimodal and temporal, containing a sequence of sound and images. We divide each input video into a set of temporally aligned images and audio segments. Compared to the synthetic scenarios, these datasets offer the complexity and noise of real-world data, thus allowing A2MT agents to optimize the trade-off objective in interesting ways.

We display examples from AudioSet in Fig. 3: (a) shows an example where both modalities are informative towards the label, in (b) the change in the image modality could be a signal for the agent to revisit the audio modality, (c) shows an example where the image modality is seemingly unrelated to the label, and in (d) all image frames are identical.

We observe empirically that model predictions degrade consistently if we mask out parts of the input at evaluation time (cf. §5). Further, we observe that models trained on inputs which span longer durations (increasing the temporal stride to keep the input size constant) perform better. These results suggest there is temporal variety in these datasets and that additional observations improve model predictions. This supports the idea that A2MT-style selection of the *right* inputs is possible.

4 Perceiver IO for A2MT

We use Perceiver IO (Jaegle et al., 2021) for both the predictive model f as well as the agent π. Perceiver IO is a suitable architecture for A2MT as it is modality-agnostic, scales well to large inputs, and supports missing input features. For the model f, we make no changes to the standard Perceiver IO architecture. For the agent π, we condition on previous actions by appending them to

the *decoder* input of the Perceiver. We propose two variants of our approach: Inputs for the synthetic datasets tend to be somewhat smaller, and we can afford a computationally more involved routine; in contrast, for the real world datasets, we use a computationally leaner setup that allows for sequential application of the Perceiver and training of the agent at scale. We give full details in §C.

4.1 Variant 1: Small Data Regime

For the synthetic scenarios, the agent π and the model f are separate Perceiver IO models that do not share any parameters and are trained jointly. For agent training, we make use of the straight-through Gumbel-Softmax estimator (Jang et al., 2017). The Gumbel trick (Jang et al., 2017; Maddison et al., 2017) allows for backpropagation through the discrete action variables and is generally considered a low-variance estimator in comparison to alternatives such as REINFORCE (Williams, 1992).

Our training directly follows Alg. 1 and Eq. (1). For a batch of training samples, we apply the agent π, unmasking modalities at timesteps according to the sampled actions a. We then predict with the model f given the partially observed input sequences $\tilde{\boldsymbol{x}}_{1:T}$. We can compute a Monte-Carlo sample of the reward, Eq. (1), given the observed prediction loss $\mathcal{L}(f(\tilde{\boldsymbol{x}}_{1:T}), y)$ and cost of acquisition $C(\mathbf{a})$, where $\mathcal{L}$ is log-likelihood loss. Due to the Gumbel parameterization of the actions, we can directly apply gradient-based optimization to the agent parameters $\boldsymbol{\theta}$. We simultaneously train the model f by minimizing the loss $\mathcal{L}$ in the objective with respect to the parameters of f.

4.2 Variant 2: Large Data Regime

We apply the Perceiver-based agent *repeatedly* during A2MT, i.e. up to $T = 25$ times in the scenarios that we study. This increases computational cost, particularly for gradient computation during training. To save compute for the large inputs of real-world datasets, we share Perceiver encoders between the predictive model f and agent π. We further pre-train the encoder and the decoder of f, and then fix both the (shared) encoder and decoder of f while we train the agent decoder parameters $\boldsymbol{\theta}$.

To ensure that the pre-trained encoder is suitable for inputs encountered later during agent training, we propose to use a particular masking setup for the inputs during pre-training. Concretely, for each input, **(M1)** we drop modalities at timestep t by sampling masks with fixed per-modality probability, $a_{m,t} \sim \text{Bernoulli}(p_m)$; **(M2)** we randomly draw $t_{\text{max}} \sim \text{Unif}(0, T)$ and mask out *all* inputs at $t > t_{\text{max}}$, i.e. $a_{m,t} = 0 \, \forall m \, \forall t > t_{\text{max}}$; **(M3)** we randomly drop entire modalities from the inputs with fixed per-modality rates $a_{m,t} = d_m \, \forall t$, with $d_m \sim \text{Bernoulli}(p_m^{(d)})$.

This masking procedure exposes the encoder to sparse input distributions during pre-training that are equivalent to those created by a randomly acting agent. This leads to versatile encoder representations that are useful for agent training. In addition to helping agent training, we observe that the masking procedures affect the test set performance of model predictions positively, and we observe the highest performance when all methods (**M1–3**) are applied simultaneously (cf. §5). This fits with similar results on how Transformer-based architectures benefit from masking at training time (Devlin et al., 2018; Dosovitskiy et al., 2021; Feichtenhofer et al., 2022).

Following related work in active feature acquisition, such as Li & Oliva (2021), we additionally condition the agent directly on model predictions $f(\tilde{\boldsymbol{x}}_{1:t})$ at each timestep t. Concretely, we concatenate both predicted class probabilities $p_f(y|\tilde{\boldsymbol{x}}_{1:t})$ as well as associated predictive entropies, $\mathbb{E}_{p_f(y|\tilde{\boldsymbol{x}}_{1:t})}[-\log p_f(y|\tilde{\boldsymbol{x}}_{1:t})]$, to the latent representation of the Perceiver IO agent. While this incurs additional computational cost from predicting with the model at each timestep, we believe the additional information will be useful in informing agent behavior. For example, this allows the agent to be aware of the uncertainty attached to model predictions at each timestep. This way, the agent could, for example, learn to stop acquiring when model predictions are sufficiently confident.

Inspired by reward shaping (Ng et al., 1999; Li & Oliva, 2021), we optionally add an intermediate reward term to the objective Eq. (1), $I = -\alpha \sum_{t=1}^{T} \left(\mathcal{L}(f(\tilde{x}_{1:t}), y) - \gamma \mathcal{L}(f(\tilde{x}_{1:t-1}), y)\right)$, where α is a hyperparameter, and γ is the discount factor. This term directly incentivizes the agent to decrease the predictive loss of of f, which helps with credit assignment, as it is otherwise hard for the agent to learn which particular acquisition actually led to loss reduction.

The Gumbel-Softmax estimator becomes prohibitively expensive in the large-scale scenario, as it requires full backpropagation through repeated application of both f and π. We therefore rely on the advantage actor critic (A2C) (Sutton & Barto, 2018) policy gradient method, which approximates gradients of Eq. (1) with respect to the agent parameters via Monte Carlo samples of the score func-

```
                | Input Sequence 1      | Input Sequence 2      | Input Sequence 3    |
----------------+-----------------------+-----------------------+---------------------
Digit           | [2 0 0 2 0 0 2 2 0 1] | [2 1 2 2 0 1 0 1 1 1] | [1 2 1 1 2 1 0 1 0 1]
Actions Digit   | [0 1 0 0 1 0 0 1 0 0] | [0 1 1 0 0 1 0 0 1 0] | [0 1 0 0 1 0 0 1 0 0]
Counter         | [1 0 2 1 0 2 1 0 2 1] | [2 1 0 2 1 0 2 1 0 2] | [1 0 2 1 0 2 1 0 2 1]
Actions Counter | [1 1 1 1 0 1 0 1 1 0] | [0 1 1 1 0 0 1 0 1 0] | [1 1 1 1 0 1 0 1 1 0]
Label           | True 2  / Pred.: 2    | True 4  / Pred.: 4    | True 5  / Pred.: 5
```

Figure 4: For the synthetic scenario, the Perceiver IO agent successfully learns a sparse acquisition strategy. 'Digit' and 'Counter' give ground truth values for the input sequence. 'Actions Digit' and 'Actions Counter' mark when in each modality the agent did (**1**) or did not acquire (**0**).

tion estimator. A2C uses an additional baseline model that reduces variance during optimization, and we implement this baseline as a separate Perceiver IO decoder head.

5 Experiments

We next give experimental results on synthetic and real data and refer to §C for further details.

5.1 Synthetic Scenario

We begin by exploring the performance of our small-scale variant on the synthetic datasets. For raw digits as inputs, our agent acquires an average of 42.2% of the digit modality, 68.8% of the counter modality, and achieves an accuracy of 92.4% on the test set. Clearly, the agent learns a selective acquisition procedure for both modalities without sacrificing predictive accuracy. We display examples of learned agent behavior on individual test set inputs in Fig. 4. For the digit modality, the agents mostly follow the ideal strategy and acquire only whenever the counter is zero. For the counter modality, the agent learns to skip acquisitions at many of the (non-informative) timesteps. In Fig. A.1, we display training dynamics: the agent initially shows high acquisition rates for both modalities, which quickly leads to increases in model predictive accuracy. Then, the agent learns to discard irrelevant parts of the input; this happens more quickly for the digit than the counter modality.

Figure A.2 shows training curves for early results on the image and image/audio versions of the synthetic scenario. The agents overfit to the training set and their behavior does not generalize. We suspect that our method struggles to learn acquisition strategies and representations simultaneously.

5.2 Audio-Visual Datasets

Next, we investigate the performance of the large scale variant of our approach on the AudioSet and Kinetics datasets. We split the training set of each dataset into a subset used for model pre-training and a subset used exclusively for agent training, taking up 80% and 20% of the original training set respectively. Note again that model and agent share the encoder, which is learned during model pre-training, and fixed during agent training.

5.2.1 Model Pretraining

Masking Variants. Table 1 gives results of different pre-training variants for Perceiver IO on AudioSet. We observe that masking significantly improves mean average precision (mAP; higher is better) on the AudioSet validation set, going from 0.178 mAP without any masking to 0.344 mAP for the complete masking setup **(M1–M3)** (cf. §4.2). We observe that each of the different masking variants **(M1-M3)** progressively improves evaluation metrics. See §C.4 for details on masking

Table 1: Model variants on AudioSet. The proposed masking mechanisms **(M1-M3)** successively improve performance.

Variant	mAP
No Masking	0.178
(M1) Random Masking	0.230
+ **(M2)** Max. Timestep	0.262
+ **(M3)** Modality Dropping	0.344
+ Conv. Downsampling	0.370

Table 2: Evaluation with missing modalities. Performance is best on fully observed data, audio is more informative for AudioSet (images for Kinetics).

Variant	AudioSet (mAP)	Kinetics (top-1)
Dense	0.370	0.396
Audio Only	0.302	0.087
Image Only	0.137	0.305

rates. We further add a single strided convolutional layer per modality, which further improves performance. Note that, here, we evaluate the model on *fully observed data*. For A2MT we are particularly interested in the predictive performance of f for sparse data.

Sparsity at Evaluation Time. In A2MT, the model f encounters sparse data during training. Table 2 reports model performance when dropping entire modalities. We observe that the model relies on both modalities for prediction, as the performance on fully observed data is best. Further, the audio modality is more informative for AudioSet and images are stronger for Kinetics. Table 3 shows how predictive quality degrades when successively dropping the stronger modality. Predictions are best on fully observed data and deteriorate significantly at 90% missing inputs. We observe that the model learns to make use of additional data in the input, and that it continues to predict reasonably for sparse inputs.

Table 3: Model performance at increasing levels of input masking. Performance degrades smoothly as the mask rate increases.

Mask Rate	AudioSet (mAP)	Kinetics (top-1)
0.0	0.370	0.396
0.3	0.369	0.397
0.8	0.331	0.365
0.9	0.292	0.316
1.0	0.137	0.087

In summary, the proposed masking routines **(M1-M3)** help learn models which use all modalities in the input and degrade as inputs become increasingly sparse. These results suggest that our pretrained Perceiver IO models are good candidates for use in A2MT, which we investigate next.

5.2.2 Agent Training.

Agents React to Cost. We follow the setup described in §4.2 and train agents using A2C without intermediate rewards. We set a nonzero acquisition cost only for the more informative modality of each dataset. Table 4 shows agent behavior for acquisition costs on the audio modality of the AudioSet dataset and Table 5 gives results for image cost on Kinetics. We observe that our agents generally react to increased acquisition costs by decreasing the number of acquisitions of that modality. We observe that agents hold on to a single acquisition (corresponding to $1/25$ frames, i.e. 0.05 acquisition rate) even at relatively high costs. This makes sense, as performance deteriorates drastically when all of the informative modality is dropped. Conversely, agents readily learn to not acquire large portions of the informative input modality, as we have observed this has only a small effect on predictive scores. For the Kinetics dataset, the agents discover that an acquisition rate of about 0.2 optimizes the objective across a variety of low to medium costs. In §A, we discuss results when imposing costs on the weak modality.

Random Ablations. We are ultimately interested in learning agents which display *adaptive* behavior that meaningfully adjusts to the peculiarities of different inputs. Therefore, we compare the performance of our agents against a random ablation that matches the average acquisition rates across modalities learned by the agent ('Random-Rate' in Table 4 and Table 5). We also explore a random agent that directly matches the average number of agent acquisitions by acquiring at a

Table 4: The agent appropriately reduces the number of acquisitions of the audio modality (acquisition rate decreases) on the **AudioSet** test set as acquisition costs increase. The agent outperforms a rate-matched ablation at high acquisition costs. However, the agent does not improve over a discrete 1-hot ablation. We give standard deviations over 5-fold repetitions over the test set.

Cost per Audio Acquisition	1×10^{-6}	1×10^{-5}	1×10^{-4}	2.5×10^{-4}	1×10^{-3}
Agent Acquisition Rate	0.98	0.83	0.23	0.11	0.04
Agent (mAP)	0.3724 ± 0.0010	0.3697 ± 0.0025	0.3359 ± 0.0117	0.3116 ± 0.0108	0.2580 ± 0.0010
Random-Rate (mAP)	0.3728 ± 0.0005	0.3719 ± 0.0007	0.3365 ± 0.0016	0.2946 ± 0.0014	0.2295 ± 0.0009
Random-1Hot (mAP)	0.3732 ± 0.0001	0.3731 ± 0.0002	0.3472 ± 0.0003	0.3165 ± 0.0006	0.2660 ± 0.0004

Table 5: The agent generally reduces its acquisition rates as costs increase for the image modality of the **Kinetics** dataset. We report standard deviations over 5-fold repeated application on the test set.

Cost per Image Acquisition	1×10^{-3}	5×10^{-3}	1×10^{-2}	1×10^{-1}	5×10^{-1}
Agent Acquisition Rate	0.19	0.22	0.21	0.09	0.04
Agent (top1)	0.3370 ± 0.0009	0.3336 ± 0.0034	0.3326 ± 0.0016	0.3309 ± 0.0011	0.2722 ± 0.0005
Random-Rate (top1)	0.2804 ± 0.0019	0.2856 ± 0.0019	0.2830 ± 0.0006	0.2378 ± 0.0012	0.1721 ± 0.0014
Random-1Hot (top1)	0.3038 ± 0.0003	0.3067 ± 0.0003	0.3063 ± 0.0002	0.2957 ± 0.0007	0.2089 ± 0.0002

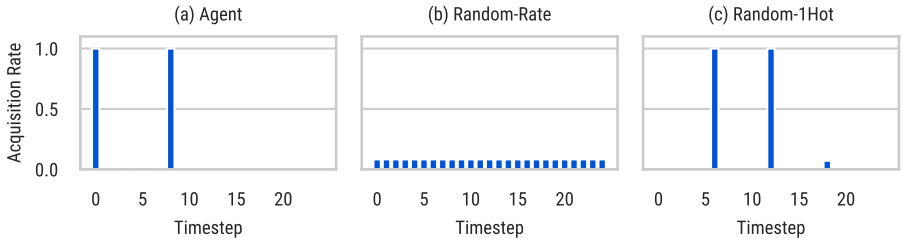

Figure 5: Comparing learned acquisition patterns of an agent on AudioSet to the patterns of the random ablations. Our agent learns a set of fixed timesteps for which it always acquires, similar to the random-1hot baseline. Small acquisition rates close to zero are too small to be visible here.

fixed set of equidistantly spaced timesteps ('Random-1Hot'). Note these are ablations rather than baselines, as they use the per-modality acquisition rates found by the agent.

For AudioSet, we observe that our agent does not consistently outperform either ablation. However, the agent does tend to outperform the random-rate, but not the random-1hot, ablation at medium-to-high acquisition costs. Figure 5 displays the average *temporal acquisition pattern* of the agent and ablations for the audio modality of AudioSet at high acquisition costs. The agent learns a *discrete* pattern of acquisitions across timesteps, similar to the random-1hot baseline. This is advantageous in comparison to the fixed low rates of the random-rate baseline: due to the small value of the acquisition probabilities, the resulting Binomial distribution over acquisitions has significant mass at 0 acquisitions. In other words, the random-rate baseline sometimes does not acquire anything at all, which has a large negative effect on its average performance.

It is encouraging that the agents learn discrete acquisition behavior and avoid the drawbacks of fixed small acquisition rates. However, this also shows that the agents do not learn *individualized* predictions, and instead acquire at fixed timesteps ignoring the inputs. There is almost no variance in acquisition behavior for different test samples. For scenarios with higher learned acquisition rates—where the variance of fixed-rate acquisition is less disadvantageous—we find some agents do not learn 1-hot style acquisition patterns and instead behave more similarly to the random-rate baseline.

For the Kinetics dataset, we observe that our agents are able to outperform both the random-rate *and* the random-1hot baseline. However, instead of individualized behavior, Figs. 6 and A.3 show that the agents learn a static pattern where they never acquire both modalities at the same time. While this is interesting behavior that optimizes the objective better than the random ablations (presumably because at least one modality is always acquired), it falls short of our goal of *adaptive* acquisition.

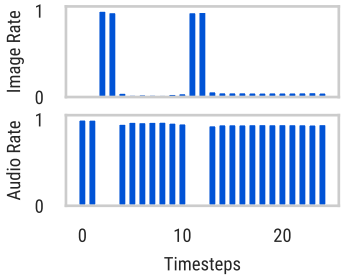

Figure 6: Learned patterns on Kinetics at 0.001 image cost.

Intermediate Rewards. We investigate the effect of intermediate rewards on agent behavior. Figure 7 (a) displays 100 random samples of agent behavior from the AudioSet test set when using intermediate rewards. We can observe that intermediate rewards generally lead to agents that show *variable* acquisition behavior. It also seems that intermediate reward—which encourages the agent to reduce predictive loss with each action—leads to greedy behavior in the agents, where acquisitions at earlier timesteps are preferred. Further, Fig. 7 (b) shows the number of acquisitions per sample is now correlated to the entropy of the predictive model. However, Fig. 7 (c) shows that model entropy and loss are practically uncorrelated here. With intermediate rewards, we find that the AudioSet agents are consistently outperformed by the random baseline, cf. Table 8. For Kinetics, we observe similar behavior (except that we outperform the random ablations, see above), cf. Table 9. We discuss this further in §7.

6 Related Work

Additional Applications of A2MT. A2MT-like problems can be found in a variety of application domains. In active perception one models the cost associated with visual attention in the context of embodied agents, e.g. Bajcsy (1988); Aloimonos et al. (1988); Mnih et al. (2014); Eslami et al. (2016); Jayaraman & Grauman (2018). Literature in AFA mentions computer security or fraud

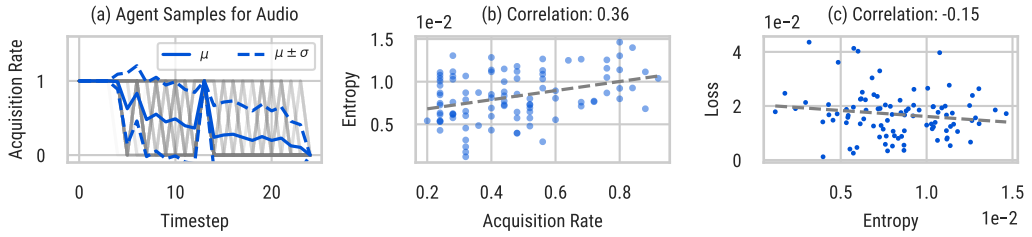

Figure 7: Behavior for agents trained with intermediate reward on 100 random samples from the AudioSet test set. (a) The learned acquisition behavior displays significant variance, and agents prefer acquisition at earlier timesteps. (Individual samples in grey, mean μ and standard deviation σ in blue.) (b) The per-sample acquisition rates are correlated to the entropy of model predictions. (c) Correlation between model entropy and predictive loss is poor.

detection as areas of application. Given that both these areas naturally may have temporal or multi-modal components, these applications transfer to A2MT as well. Lastly, the acquisition procedure in A2MT can be a way to reduce input size and computational cost of predictions of the model f.

Efficient Video Classification. Related work in efficient video classification has sought to reduce the computational cost of video classification by selecting a small set of salient frames for each input (Yeung et al., 2016; Wu et al., 2019b; Korbar et al., 2019; Wu et al., 2019a; Gao et al., 2020; Zheng et al., 2020; Wang et al., 2021; Ghodrati et al., 2021; Panda et al., 2021; Gowda et al., 2021; Yang et al., 2022). For computationally cheap policies, these methods can achieve computational cost-savings compared to methods that rely on all timesteps as input. In contrast, A2MT assumes that there is a real-world cost associated with the *acquisition* of the input modalities, e.g. the cost of performing an MRI scan, and we can completely neglect any computational costs. Also, A2MT requires temporally causal acquisition, which is something not respected by these approaches. These difference in motivation lead to methods which (in all cases we know of) are not applicable to A2MT.

7 Discussion

The results presented in Fig. 7 (§5.2.2) raise the question if the models we use are sufficiently powerful to successfully learn adaptive behavior. From an information-theoretic perspective the agent behavior, acquiring proportionally to the entropy, cf. Fig. 7 (b), is sensible: for samples where the model is certain about the prediction (low entropy) the agent can stop acquisitions early; when the model is uncertain, the agent may acquire more to reduce uncertainty. However, for this to make sense, reductions in predictive entropy should correspond to reductions in actual loss incurred. Figure 7 (c) shows this is not the case in our experiments. This suggests there might not be enough signal from the model predictions for the agent to anticipate how acquisitions affect the reward.

For future work, we would expect improved model uncertainty to help with A2MT performance. Alternatively, subsequent work in A2MT could consider datasets with longer clip durations, e.g. ActivityNet (Fabian Caba Heilbron & Niebles, 2015): while we have found that there is some signal diversity in AudioSet and Kinetics, different sections of a given clip often look similar, making it hard for the agents to learn adaptive behavior. There are also a variety of interesting variations of the A2MT setup, e.g. other reward formulations such as a fixed budget per sample or a global budget across samples, acquisition costs that change with time, or actions that affect state evolution.

8 Conclusion

We have introduced active acquisition for multimodal temporal data (A2MT), as well as a Perceiver IO–based reinforcement learning approach to tackle A2MT problems. We introduce synthetic scenarios and show our agents successfully learn to use cross-modal reasoning. On real world data, our agents appropriately react to modality-specific acquisition costs. However, ablations reveal they are unable to successfully learn individualized behavior. We believe A2MT is a challenging and practically relevant task, and we would be excited for the community to join our efforts.

ACKNOWLEDGMENTS

We thank Yujia Li, Timothy Lillicrap, Andrew Brock, Adrià Recasens, João Carreira, Lucas Smaira, Isabela Albuquerque, Joost van Amersfoort, Ali Eslami, and the anonymous reviewers for helpful feedback and interesting discussions that have led to numerous improvements of the paper.

REFERENCES

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4):333–356, 1988.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pp. 273–304, 1995.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Andres Diaz-Pinto, Nishant Ravikumar, Rahman Attar, Avan Suinesiaputra, Yitian Zhao, Eylem Levelt, Erica Dall'Armellina, Marco Lorenzi, Qingyu Chen, Tiarnan DL Keenan, et al. Predicting myocardial infarction through retinal scans and minimal personal information. *Nature Machine Intelligence*, 4(1):55–61, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in Neural Information Processing Systems*, 29, 2016.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv:2205.09113*, 2022.

Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Conference on Computer Vision and Pattern Recognition*, pp. 10457–10467, 2020.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 15608–15618, 2021.

Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 1451–1459, 2021.

Russell Greiner, Adam J Grove, and Dan Roth. Learning cost-sensitive active classifiers. *Artificial Intelligence*, 139(2):137–174, 2002.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

Z. Jackson. Free spoken digit dataset (fsdd). https://github.com/Jakobovski/free-spoken-digit-dataset, 2017.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. In *International conference on machine learning*, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 3959–3966, 2019.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features as a sequential decision-making problem. *Machine Learning*, 109(8):1587–1615, 2020.

Dinesh Jayaraman and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *Conference on computer vision and pattern recognition*, pp. 1238–1247, 2018.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv (version 2.0). *PhysioNet*, 2020.

Mohammad Kachuee, Orpaz Goldstein, Kimmo Karkkainen, Sajad Darabi, and Majid Sarrafzadeh. Opportunistic learning: Budgeted cost-sensitive learning from data streams. In *International Conference on Learning Representations*, 2019.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *International Conference on Computer Vision*, pp. 6232–6242, 2019.

Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Sarah Lewis, Tatiana Matejovicova, Yingzhen Li, Angus Lamb, Yordan Zaykov, Miltiadis Allamanis, and Cheng Zhang. Accurate imputation and efficient data acquisitionwith transformer-based vaes. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *International Conference on Machine Learning*, pp. 6450–6459. PMLR, 2021.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Charles X Ling, Qiang Yang, Jianning Wang, and Shichao Zhang. Decision trees with minimal costs. In *International conference on Machine learning*, pp. 69, 2004.

Lance A Liotta, Mauro Ferrari, and Emanuel Petricoin. Clinical proteomics: written in blood. *Nature*, 425(6961):905–905, 2003.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv:1809.11142*, 2018.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pp. 483–486. IEEE, 2004.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video recognition. In *International Conference on Computer Vision*, pp. 7576–7585, 2021.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009.

Paola Sebastiani and Henry P Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1): 145–157, 2000.

Victor S Sheng and Charles X Ling. Feature value acquisition in testing: a sequential batch test algorithm. In *International conference on Machine learning*, pp. 809–816, 2006.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. *Advances in neural information processing systems*, 31, 2018.

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv:2010.10864*, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *International Conference on Computer Vision*, pp. 16249–16258, 2021.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *International Conference on Computer Vision*, pp. 6222–6231, 2019a.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 1278–1287, 2019b.

Mingyu Yang, Yu Chen, and Hun-Seok Kim. Efficient deep visual and inertial odometry with adaptive visual modality selection. *arXiv:2205.06187*, 2022.

Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *computer vision and pattern recognition*, pp. 2678–2687, 2016.

Haiyan Yin, Yingzhen Li, Sinno Jialin Pan, Cheng Zhang, and Sebastian Tschiatschek. Reinforcement learning with efficient active feature acquisition. *arXiv preprint arXiv:2011.00825*, 2020.

Sara Zannone, José Miguel Hernández-Lobato, Cheng Zhang, and Konstantina Palla. Odin: Optimal discovery of high-value information using model-based deep reinforcement learning. In *ICML Real-world Sequential Decision Making Workshop*, 2019.

Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. *Transactions on Image Processing*, 29:7970–7983, 2020.

Valentina Bayer Zubek, Thomas Glen Dietterich, et al. Pruning improves heuristic search for cost-sensitive learning. 2004.

A Additional Results

In Fig. A.1, we display the training dynamics for the experiments on the synthetic scenario with raw inputs, Fig. A.2 gives performance for the MNIST and the MNIST + Spoken Digit version of the synthetic experiment.

Table 6 and Table 7 give results of our agents on AudioSet and Kinetics respectively when costs are imposed on the weaker input modality. For AudioSet, the agents learn to acquire $\approx 1/25$ image per sequence across a magnitude of acquisition costs. Surprisingly, the performance at 1 image frame is only about 1 p.p. mAP lower than the performance on fully observed data, which explains why it makes sense for the agent to quickly drop to such low acquisition rate on the image modality. At high costs of 0.01 per acquisition, the agent no longer acquires any images, which finally does hurt mAP by about 4 p.p. For Kinetics, we observe that the number of acquisitions for the weaker audio modality drops steadily as cost is increased, and, correspondingly, so does the top1-accuracy. This indicates that, for Kinetics, the model makes better use of the weaker audio modality.

Figure A.3 shows how the agent learns to never acquire both modalities on the Kinetics dataset across a variety of costs.

Table 8 and Table 9 give agent behavior with intermediate rewards enabled for AudioSet and Kinetics. This do not lead to significantly improved agent performance.

B Synthetic Sequence Generation

Algorithm 2 gives Python code for generating the sequences for the synthetic dataset (cf. §3.1).

Algorithm 2 Generate Synthetic Sequence

```
def draw_sequence(length, digit_low, digit_high, counter_low, counter_high):

  digits, counter, important_digits = [], [], []

  while len(digits) < length:
    max_count = np.random.randint(counter_low + 1, counter_high + 1)
    counter_i = np.arange(max_count, counter_low - 1, -1)
    counter.extend(counter_i)

    digits_i = np.random.randint(digit_low, digit_high + 1, len(counter_i))
    if len(digits_i) + len(digits) <= length:
      important_digits.append(digits_i[-1])
    digits.extend(digits_i)

  digits = digits[:length]
  counter = counter[:length]
  label = sum(important_digits)

  return digits, counter, label
```

Table 6: Acquisition behavior for the agent and ablations as costs increase for the image modality of the AudioSet dataset. Standard deviations over 5 repetitions over the test set.

Cost Image	0.00001	0.00010	0.00100
Agent Acquisition Rate	0.07	0.05	0.01
Agent (mAP)	0.3683 ± 0.0060	0.3627 ± 0.0017	0.3289 ± 0.0072
Random-Rate (mAP)	0.3505 ± 0.0005	0.3470 ± 0.0010	0.3238 ± 0.0008
Random-1Hot (mAP)	0.3681 ± 0.0003	0.3650 ± 0.0004	0.3275 ± 0.0004

Table 7: Acquisition behavior for the agent and ablations as costs increase for the audio modality of the Kinetics dataset. Standard deviations over 5 repetitions over the test set.

Cost Audio	0.001	0.010	0.100
Agent Acquisition Rate	0.79	0.43	0.05
Agent (top1)	0.3313 ± 0.0041	0.3253 ± 0.0036	0.2672 ± 0.0015
Random-Rate (top1)	0.2825 ± 0.0012	0.2840 ± 0.0020	0.2524 ± 0.0013
Random-1Hot (top1)	0.3063 ± 0.0003	0.2930 ± 0.0003	0.2638 ± 0.0004

Table 8: Agent behavior with intermediate rewards on AudioSet with different audio acquisition costs and trade-off parameters α. See §4.2 for an explanation of intermediate rewards and α. Standard deviations over 5 repetitions over the test set.

Cost Audio	0.0005	0.0005	0.00010	0.00010
Trade-off α	0.1	0.5	0.1	0.5
Agent Acquisition Rate	0.05	0.13	0.04	0.06
Agent (mAP)	0.2648 ± 0.0063	0.3104 ± 0.0120	0.2671 ± 0.0009	0.2668 ± 0.0075
Random-Rate (mAP)	0.2439 ± 0.0012	0.3043 ± 0.0013	0.2323 ± 0.0014	0.2570 ± 0.0014
Random-1Hot (mAP)	0.2750 ± 0.0006	0.3238 ± 0.0004	0.2664 ± 0.0003	0.2838 ± 0.0008

Table 9: Agent behavior with intermediate rewards on Kinetics with different image acquisition costs and trade-off parameters α. See §4.2 for an explanation of intermediate rewards and α. Standard deviations over 5 repetitions over the test set.

Cost Image	0.01	0.01	0.10	0.10
Trade-off α	0.1	0.5	0.1	0.5
Agent Acquisition Rate	0.19	0.27	0.15	0.18
Agent (top1)	0.3260 ± 0.0018	0.3234 ± 0.0012	0.3225 ± 0.0027	0.3162 ± 0.0008
Random-Rate (top1)	0.2895 ± 0.0004	0.2673 ± 0.0011	0.2768 ± 0.0011	0.2807 ± 0.0009
Random-1Hot (top1)	0.3015 ± 0.0003	0.2798 ± 0.0004	0.3042 ± 0.0004	0.3037 ± 0.0003

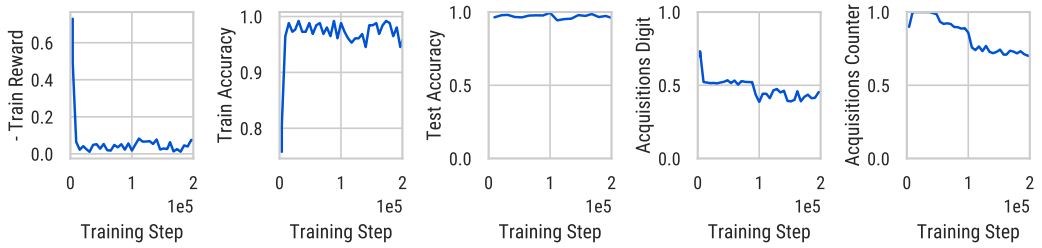

Figure A.1: Performance of the Perceiver IO-based model and agent when applied to the raw-digit version of the synthetic dataset. The model learns to quickly solve the task, and the agent slowly reduces the number of acquired datapoints. Note that the 'Acquisitions' plots give acquisition rates s.t. a '1' corresponds to the entire modality being acquired.

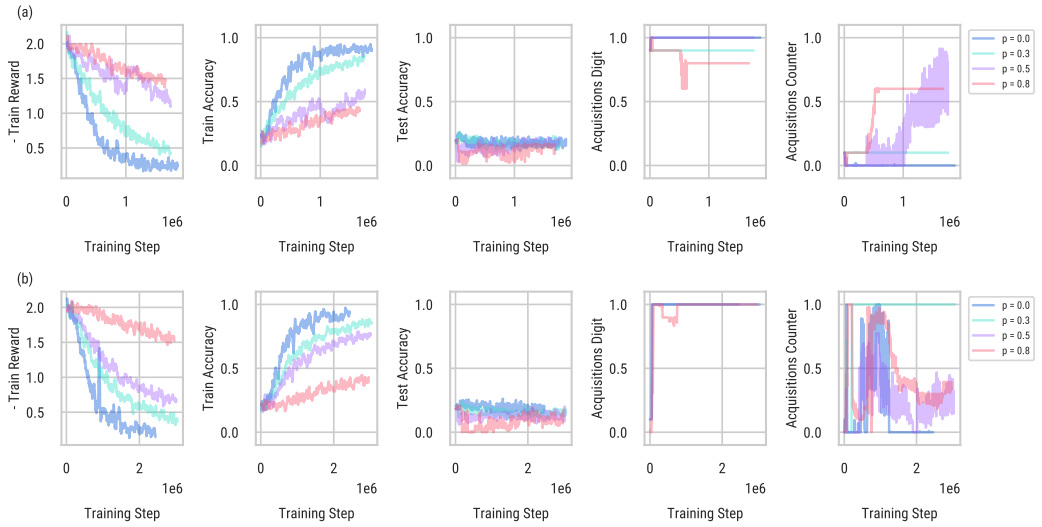

Figure A.2: Performance of the Perceiver IO-based model when applied to the (a) MNIST and (b) MNIST/Spoken Digit version of the synthetic dataset. We are unable to learn prediction strategies that generalize to the test set for this scenario. Different plots show sweep over dropout probabilities in $[0, 0.3, 0.5, 0.8]$ for the input tokens: while we do observe a regularizing effect for dropout here, it does not lead to improved generalization performance. Acquisition cost is set to 0 for these runs. Note that the 'Acquisitions' plots give acquisition rates s.t. a '1' corresponds to the entire modality being acquired. We apply running mean smoothing with kernel size 10 for the plots of the first two columns.

C Details on Synthetic Experiments

C.1 Synthetic Experiments: Raw Inputs

C.1.1 Dataset

We generate a synthetic dataset with training set of size 50×10^3 and test set of size 10×10^3. For the 'raw' version, the input modalities have shape $(10, 3)$, i.e. we use sequences of length 10 and there are three distinct values $(2, 1, 0)$ per modality. We set the acquisition cost to 0.0005 per modality and timestep.

C.1.2 Architecture

We train using a batch size of 256. We use the ADAM optimizer with initial learning rate of 3×10^{-4}, weight decay of 1×10^{-6}, and a cosine annealing schedule. For the Perceiver IO encoder we use

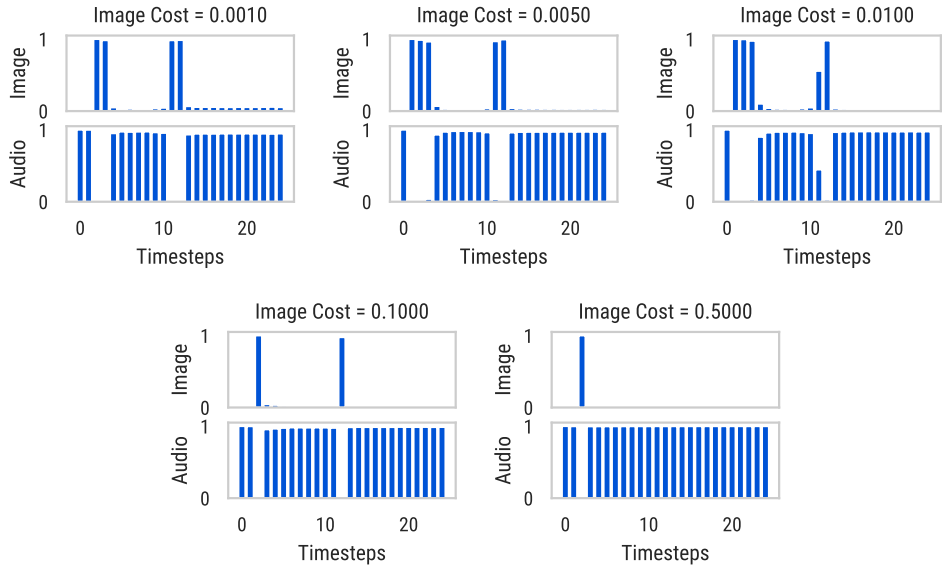

Figure A.3: Learned patterns on Kinetics at a variety of image cost. The y-axis gives the average acquisition rate of the agent for a modality and timestep. As the image cost increases, the number of acquisitions decreases. Further, the agents learn a static pattern whereby at least one modality is always acquired.

a single cross-attend block with 4 self-attention operations per Perceiver IO block, and we use 128 queries, and the hidden dimension is 128. For the Perceiver IO decoder, we use a single head with 128 queries and hidden dim of 128. We train for a total of 2×10^5 steps. We set the discount factor to $\gamma = 1$.

C.2 Synthetic Experiments: MNIST and SpokenDigit Versions

For all parameters not mentioned in the below, we use the same settings as for the 'raw' version of the synthetic dataset. We use a batch size of 128 and train for $> 1 \times 10^6$ steps. The input shape of the MNIST images is $(10, 28, 28)$, and the shape of the audio snippets is $(10, 39, 80)$ after mel spectrogram pre-processing.

C.3 Audio-Visual Datasets

C.3.1 Dataset

The AudioSet dataset has a training set of size 1 771 873, an evaluation set of size 17 748, and a total of 632 classes. We use the unbalanced version of the dataset. The Kinetics dataset has a training set of size 545 793 a test set of size 67 858, and 700 classes. For both datasets, each input sample consists of audio-visual input with 250 frames of images and a raw audio signal spanning a total of 10 seconds. For images, we take every 10th frame as input, obtaining a total of 25 input frames. We pre-process the audio signal to mel spectrograms, and then divide the signal into 25 segments. The audio modality has input shape $(25, 40, 128)$, and the image modality has input shape $(25, 200, 200, 3)$.

C.4 Architecture.

Perceiver IO. For the Perceiver IO encoder we use a single cross-attend block with 8 self-attention operations per Perceiver IO block, and we use 1024 queries, and the hidden dimension is 1024. For the Perceiver IO decoder, i.e. the policy head, we use a single cross-attend head with 1024 queries and hidden dim of 1024. The A2C baseline network is also a Perceiver IO decoder module with the same configuration as for the policy.

Pre-Training. We use the ADAM optimizer with initial learning rate of 3×10^{-4}, weight decay of 1×10^{-6}, and a cosine annealing schedule. We train for a total of 100 epochs using a batch size of 512. We use the masking settings detailed below.

For the masking variants **M1-M3**, we report results with the following settings. For **M1**, we drop inputs at given modality and timestep with $p_m = 0.8$. For **M2**, we set $p_m = 0.6$, which accounting for the additional 'max-timestep' masking mechanism, yields the same expected masking rate of 0.8. For **M3**, we keep $p_m = 0.6$ and additionally drop modalities with $d_m = 0.5$. In all of the above, probabilities are the same across modalities m. These settings performed best in preliminary experiments.

Agent Training. We train agents for a total of 20×10^3 steps. We re-use the optimizer configuration from pre-training for training of the agent and the A2C baseline network. We set the discount factor to $\gamma = 1$.