

# A Framework for Multisensory Foresight for Embodied Agents

Xiaohui Chen<sup>§</sup>, Ramtin Hosseini<sup>§</sup>, Karen Panetta, and Jivko Sinapov

Tufts University, MA, USA

{Xiaohui.Chen, Ramtin.Hosseini, Jivko.Sinapov}@tufts.edu, Karen@ece.tufts.edu

**Abstract**—Predicting future sensory states is crucial for learning agents such as robots, drones, and autonomous vehicles. In this paper, we couple multiple sensory modalities with exploratory actions and propose a predictive neural network architecture to address this problem. Most existing approaches rely on large, manually annotated datasets, or only use visual data as a single modality. In contrast, the unsupervised method presented here uses multi-modal perceptions for predicting future visual frames. As a result, the proposed model is more comprehensive and can better capture the spatio-temporal dynamics of the environment, leading to more accurate visual frame prediction. The other novelty of our framework is the use of sub-networks dedicated to anticipating future haptic, audio, and tactile signals. The framework was tested and validated with a dataset containing 4 sensory modalities (vision, haptic, audio, and tactile) on a humanoid robot performing 9 behaviors multiple times on a large set of objects. While the visual information is the dominant modality, utilizing the additional non-visual modalities improves the accuracy of predictions.

## I. INTRODUCTION

For humans and many animals, the ability to anticipate the future is a prerequisite for intelligent behavior. For robots, predicting the future values of sensors can assist object manipulation (*e.g.* planning towards a desired sensory state), anomaly and failure detection (*e.g.* by comparing predictions to observed values), and sensorimotor learning (*e.g.* learning how sensors change as a result of the robot’s actions). More generally, if a robot can predict the future values of sensors such as its cameras or haptic sensors, any perceptual routine that is used to process the robot’s current sensory state would also be applicable for predicted sensory states.

Early work in robotics focused on learning visual forward models that anticipate the future trajectories of objects manipulated by the robot as well as movements by the robot itself [1]. More recently, methods have been developed to directly predict the future raw image frames that the robot would observe in its camera stream over the course of object manipulation [2]. One limitation of existing methods is that they mostly deal solely in the visual domain. For many object manipulation tasks, however, other sensory modalities, such as haptic, audio and tactile, may be just as important. Non-visual sensory modalities can also help in situations where vision alone may be insufficient to resolve an ambiguity (*e.g.* two objects may look identical but one may be much heavier than the other). Indeed research conducted in cognitive science [3], [4] and robotics [5], [6] has demonstrated the importance of using multiple (and often, non-visual) sensory modalities when learning about object properties and affordances.

Motivated by these findings, we present a deep learning methodology for *multisensory foresight* which uses feedback from multiple sensory modalities produced over the course of the robot’s interaction with objects in its environment. We hypothesize that including more modalities can substantially improve prediction performance. To present and evaluate our proposed methodology, we used a publicly available dataset [7], in which a robot performed 9 different types of exploratory behaviors (*e.g.* *push*, *press*, *etc.*) on 100 objects multiple times. The dataset includes vision, haptic, audio, and vibrotactile sensory modalities. This paper introduces a modular deep neural network architecture that can take advantage of any modalities for performing the next-frame prediction task. Furthermore, we extend the model to predict the next frame for modalities other than vision, which leads to further improvements in the robot’s prediction performance.

## II. RELATED WORK

**Multi-modal perception.** A large volume of research has shown that perception can benefit by relating information from multiple sources [8], [9], [6], [10], [11]. To identify the semantics of objects (*e.g.* empty, soft), visual information alone may not be adequate as the objects could be identical in the visual domain but different in other aspects (*e.g.* material, internal state, compliance). To address this problem, several lines of research have focused on how robots can use non-visual sensory modalities of tasks that include grasping [12], [13], object recognition [14], [15], [16], object categorization [7], [17], [18] and language grounding [19], [20], [21], [22]. Inspired by these works, we propose an architecture that also uses multiple sensory modalities for the sensorimotor learning task of visual next-frame prediction.

**Frame prediction.** This research aims to forecast future frames in video sequences. Early studies have focused on employing complex networks to directly generate pixel values (*e.g.* [23]). However, these methods generally produce blurry predictions, as it is hard to model the distribution of image pixels, especially multiple steps into the future. Inspired by language modeling, Ranzato *et al.* [24] applied a recurrent neural network to anticipate future frames. Srivastava *et al.* [25] adapted LSTM model to capture pixel dynamics. Mathieu *et al.* [26] investigated different loss functions for sharper frame predictions. In another effort, Oh *et al.* [27] proposed an action-conditional autoencoder network for Atari Games. Liang *et al.* [28] defined a dual motion Generative Adversarial Net (GAN). Recently a few approaches have solved the issue of blurriness of predictions multiple steps into the future [29], [30], [31]. Despite the remarkable success, they have

<sup>§</sup>The first two authors contribute equally.

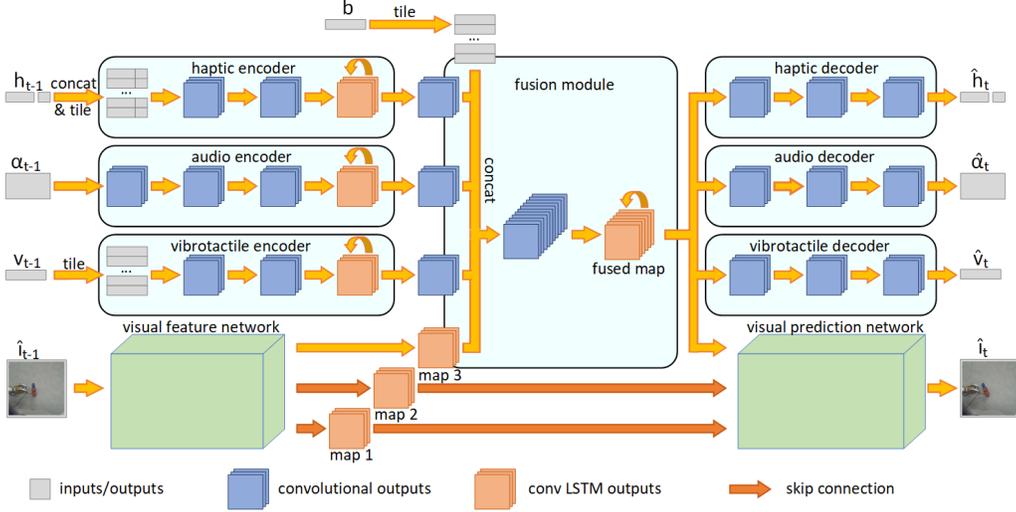


Fig. 1: The architecture of the proposed model, which consists of 4 feature encoders (left) and prediction heads (right) for 4 modalities, and 1 fusion module (middle) for merging representations of different modalities.

their own limitations. For example, [29] uses a hierarchical method which enables it to make sharper images for a longer period of time; however, it has the limitation that occasionally predictions disappear which constrains its applicability in safety critical settings. Two of the most successful models for frame prediction are PredNet [32], and the work introduced in [2]. ConvLSTM units are essential building blocks of these two models. PredNet makes local predictions in each layer of the network and only passes deviations from the predictions to succeeding layers. The model presented in [2] uses a pixel transformation function such as convolutional dynamic neural advection (CDNA) to predict motion distribution for the objects in videos. Despite the immense success, this model considers only one modality (vision) alongside state and action for forecasting future frames. In this paper, the proposed multi-modal network draws on the model architecture from [2] for the vision prediction branch. By integrating several modalities to the network, the proposed model shows significant improvement in performance compared to the single-modality network.

### III. LEARNING METHODOLOGY

Next, we describe our framework for multisensory foresight, which uses multiple sensory modalities coupled with exploratory actions performed on objects by the robot.

#### A. Notation and problem formulation

We used a dataset which contains  $N$  samples  $\{X^n\}$ , where  $n = 1, 2, \dots, N$ , and each sample  $X^n$  is defined as a quadruple  $X^n = \{\mathcal{I}^n, \mathcal{H}^n, \mathcal{A}^n, \mathcal{V}^n\}$ . The quadruple is consisted of 4 kinds of sequential data collected by different sensors: 1) Visual data  $\mathcal{I}^n = \{i_1^n, i_2^n, \dots, i_T^n\}$ ; 2) Haptic data  $\mathcal{H}^n = \{h_1^n, h_2^n, \dots, h_T^n\}$ ; 3) Auditory data  $\mathcal{A}^n = \{a_1^n, a_2^n, \dots, a_T^n\}$ ; and 4) Vibrotactile data  $\mathcal{V}^n = \{v_1^n, v_2^n, \dots, v_T^n\}$ . Different sensors of the robot execute at different frequency rate. As a

result, with regard to our primary task which is to predict the following visual frames, all other modalities are processed to be synchronized to the visual data in terms of time step. To meet this end, for each time step, the modality data is defined as follows:

$$i_t^n \in \mathbb{R}^{M_w \times M_h \times M_c}, h_t^n \in \mathbb{R}^{M_d \times M'_d}$$

$$a_t^n \in \mathbb{R}^{M_e \times M'_e}, v_t^n \in \mathbb{R}^{M_f \times M'_f}$$

where  $M_w, M_h$  and  $M_c$  are the width, height and the number of channels of each image respectively,  $M_d$  is the number of robot joint-torque sensor readings,  $M_e$  is the number of frequency bins in the audio spectrogram, and  $M_f$  is the number of accelerometer readings. Moreover,  $M'_d, M'_e$  and  $M'_f$  are the number of in-frame time steps of haptic, auditory, and vibrotactile modalities respectively.

The goal of the framework is to predict the future visual frames  $\mathcal{I}_{>K}^n = \{i_{K+1}^n, \dots, i_T^n\}$  given  $K$  context frames  $\mathcal{I}_{\leq K}^n = \{i_1^n, i_2^n, \dots, i_K^n\}$  along with other modalities, where  $K < T$ . We also add a categorical feature  $b^n \in \mathcal{B}$  indicating the type of behavior performed by the robot. While the main task is predicting subsequent frame images  $\hat{i}_t^n$ , where  $t \in \{K+1, \dots, T\}$ , we introduce the concept of auxiliary tasks learning, which also predicts the next frames for haptic, audio and vibrotactile modalities which are denoted as  $\hat{h}_t^n, \hat{a}_t^n$  and  $\hat{v}_t^n$  respectively. Auxiliary tasks are expected to help find a stronger representation of how the modalities relate to one another through backpropagation, from which the main task might benefit. To this end, we define a highly abstracted autoregressive model  $\mathcal{F}$ :

$$\hat{i}_t^n, \hat{h}_t^n, \hat{a}_t^n, \hat{v}_t^n = \mathcal{F}(\mathcal{I}_{\leq K}^n, \mathcal{H}_{< t}^n, \mathcal{A}_{< t}^n, \mathcal{V}_{< t}^n, b^n) \quad (1)$$

where  $\mathcal{H}_{< t}^n, \mathcal{A}_{< t}^n, \mathcal{V}_{< t}^n$  are the additional modality sequences prior to time step  $t$ . The model first learns how to extract

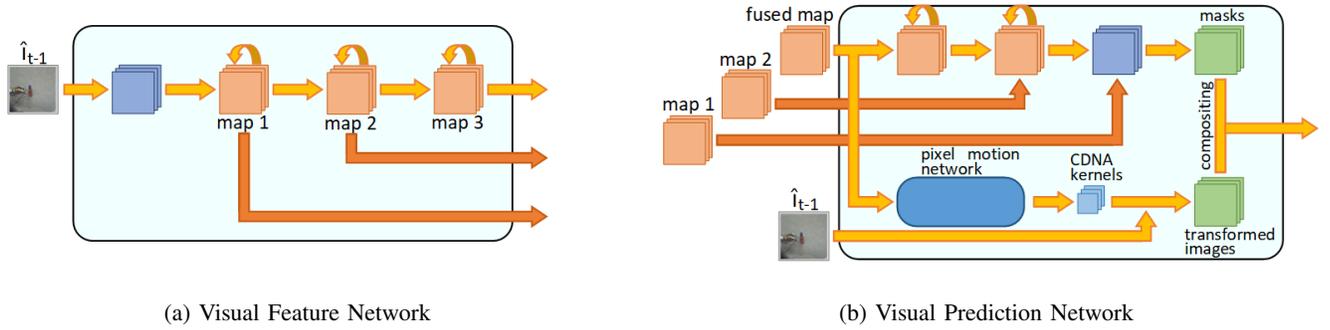


Fig. 2: Pipeline of The Visual Prediction Module, 2a shows the architecture of visual feature extractor, and 2b shows the architecture of visual prediction network.

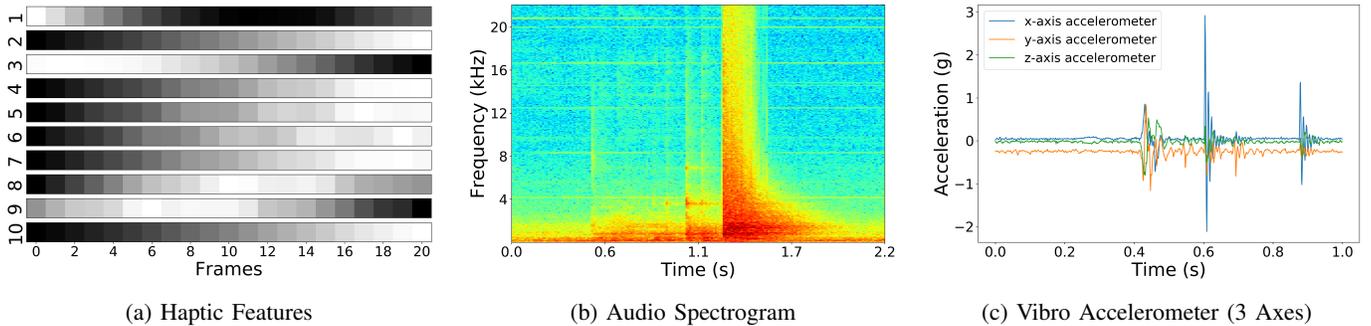


Fig. 3: Visualization of (a) haptic, (b) audio and (c) vibrotactile modalities when the robot *drops* a bottle

high-level representations of each modality individually, then learns the interaction and combination of the 4 modality representations, and finally outputs the next frame prediction of each modality using the *multi-heads* network. Next, we discuss the details about the model  $\mathcal{F}$ .

### B. Model architecture

The proposed network architecture, shown in Figure 1, consists of 3 sub-modules: feature encoders, fusion module, and multi-modality prediction network.

*a) Feature Encoders:* Previous methods on next frame prediction relied mainly on the visual modality, while in our approach, inputs to the network are sequences of different modalities  $\mathcal{I}, \mathcal{H}, \mathcal{A}, \mathcal{V}$ . To efficiently integrate different modality features together, all modalities are mapped into  $W \times H$  feature maps with different numbers of channels via their corresponding feature encoder. The feature encoder networks are composed of convolution, downsampling, and ConvLSTM modules with concatenation and tile operation.

For the visual modality, we employ stack ConvLSTMs (Figure 2a) to extract high-level vision features as well as spatio-temporal features. For the haptic modality, we spatially tile the concatenated joint signals and robot gripper pose across the feature map and feed it into the haptic-specific feature extraction network. For the audio and vibrotactile modalities, first we use Fast Fourier Transform (FFT) to compute a spectrogram, then employ convolutional layers and ConvLSTM layer to extract features.

*b) Fusion Module:* The fusion module contains one convolutional layer and one ConvLSTM layer with a concatenation operation, as illustrated in Figure 1. To further merge the modality features, it first integrates the lowest-dimensional activation maps given by each feature encoder into one latent feature map along the channel via concatenation operator, and feed it into the defined layers sequentially. The number of channels in the output feature map will be compressed into the same as of the visual input feature map, which in our work, channel size 64, and 128 are considered. The output feature map contains information extracted from all used modalities and will be further used to predict each modality in the next frame. Note that the number of chosen modalities can vary from 1 to 4, and the fusion module will automatically adapt the modality setting and output the integrated feature map with a fixed number of channels.

*c) Multi-modal prediction head:* The core of the model is learning the internal relation across different modalities, which consequently leads to increasing the performance of the main task (visual next-frame prediction). This is achieved by augmenting the auxiliary tasks. For each modality, there is a head that gets its input (fused feature map) from the fusion module, which integrates all the information and outputs the corresponding next frame modality.

For auxiliary task prediction heads, we directly reconstruct the next frame. Transposed convolutional layers are employed in each decoder, and the fused map is upsampled to be in the

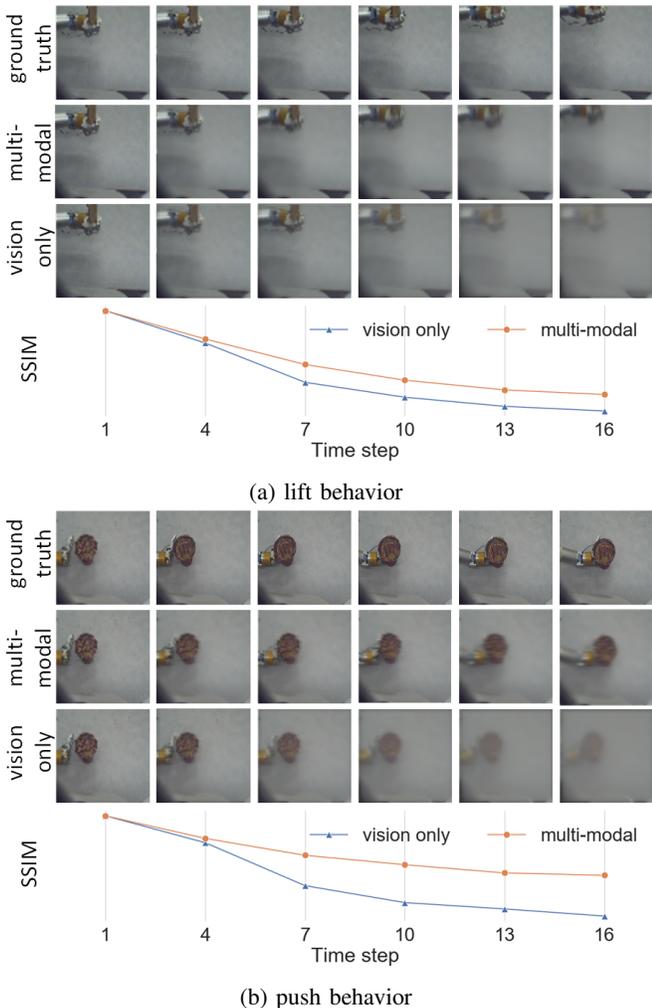


Fig. 4: Sharpness of predicted images, when the robot arm perform different behaviors (4a: *lift*, 4b: *push*).

same dimension as the original input. For the visual prediction head (Figure 2b), we use the idea of pixel transformation proposed in [2], [33], and perform two tasks instead of reconstructing the image directly. The first task is learning the pixel transformation parameters for each grouped object. The second task is performing an instance segmentation task that aims to group pixels by object. There are two branches for the visual prediction head. In the **object motion capture branch**, a motion prediction module called convolutional dynamic neural advection (CDNA) is used [2]. The CDNA function computes new pixel values by applying multiple normalized convolution kernels to previous frames. CDNA is an object-centric motion prediction module, and as it is indicated in [2], the intuition behind it is that pixels form the same rigid entity move together. This module is expressed in the following equation:

$$\hat{J}_t(x, y) = \sum_{k \in (-k, k)} \sum_{l \in (-k, k)} \hat{m}(k, l) \hat{I}_{t-1}(x - k, y - l) \quad (2)$$

where  $k$  is the size of  $\hat{m}$  convolution kernel, and  $\hat{J}$  is a set of several transformations of the previous image. In the **instance segmentation branch**, skip connections are used to include the intermediate feature maps obtained from the visual encoder to the middle of the prediction head by directly concatenating them to restore the details learned in the low-level feature maps. This branch is responsible for applying masks to different objects. Finally, to obtain a single output image  $\hat{I}_t$ , the composition of predicted images should be modulated by a mask.

$$\hat{I}_t = \sum_c \hat{J}_t^{(c)} \odot \Xi \quad (3)$$

where  $c$  represents the channel of the mask, and  $\odot$  is the Hadamard product. The total loss function contains 4 components, each of which corresponds to the cost function for each modality. The cost function for each modality is weighted and is described below.  $\mathcal{L}_T$  is the total loss:

$$\mathcal{L}_T = \lambda_i \mathcal{L}_i + \lambda_h \mathcal{L}_h + \lambda_a \mathcal{L}_a + \lambda_v \mathcal{L}_v \quad (4)$$

where in our work, the coefficient hyper-parameters are selected via grid search:  $\lambda_i = 1.0$ ,  $\lambda_h = 10^{-4}$ ,  $\lambda_a = 10^{-3}$  and  $\lambda_v = 10^{-4}$ . We used mean square error (MSE) as the cost function for each modality.

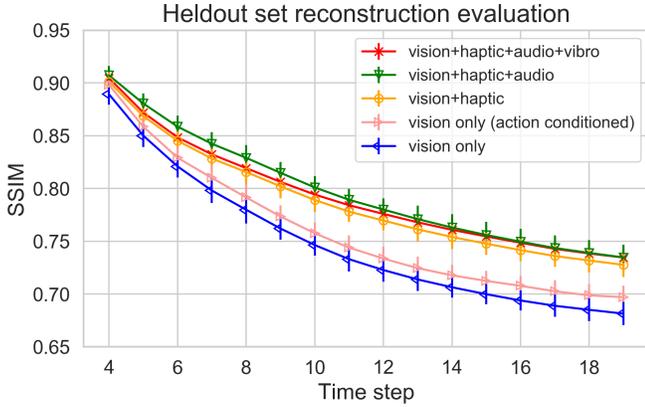
#### IV. EXPERIMENTAL RESULTS

We compare the proposed framework with the vision only model proposed in [2] both quantitatively and qualitatively. To better investigate the robustness of the model, we provide two settings for experiments, which will be discussed in sections IV-A and IV-B. Furthermore, we discuss the effect of employing auxiliary training in section IV-C.

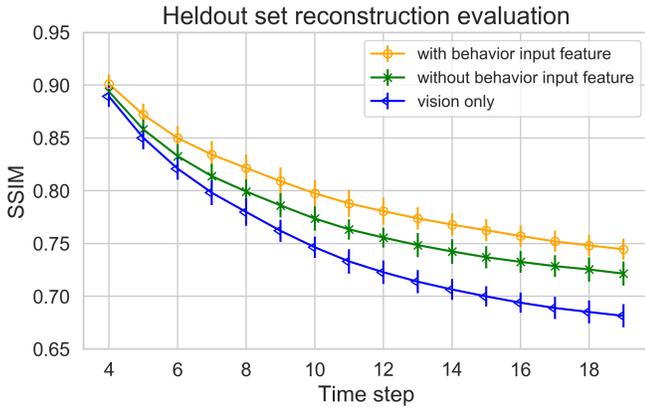
**Implementation Details.** We make use of PyTorch [34] for GPU-based implementation<sup>1</sup>, set the number of context frames  $K$  to 4, and evaluate the model performance for the following 16 predicted frames. For a few behaviors (*grasp* and *tap*), there are fewer frames in the dataset, only the following 6 frames are predicted. We employed ADAM optimizer [35] with learning rate  $lr = 1e^{-3}$  to train the network for 30 epochs with batch size 32. For evaluation, we use Structural Similarity Index Measurement (SSIM) metrics to measure the visual prediction quality. Alternative metrics, such as Maximum Mean Discrepancy (MMD) [36] could be considered. We performed 5-fold cross-validation such that during each test, data from 80 objects was used for training and data from the remaining 20 objects was used for testing.

**Dataset.** The dataset described in [37] is used to evaluate and compare the proposed network with the single-modal network. For collecting the dataset, an uppertorso humanoid robot with a 7-DOF arm manipulates 100 objects by executing 9 different exploratory behaviors (*push*, *poke*, *press*, *shake*, *lift*, *drop*, *grasp*, *tap* and *hold*) multiple times and records visual, haptic, auditory and vibrotactile sensory data. The

<sup>1</sup>Code: <https://github.com/tufts-ai-robotics-group/mmvp/tree/main>



(a) Ablation study on sensory inputs



(b) Ablation study on behavior input

Fig. 5: Quantitative result evaluated with SSIM metric. Ablation studies on all behavior setting

TABLE I: Investigation of contribution of each modality to the improvement of model prediction

avg. SSIM	haptic	audio	vibrotactile	behavior
0.771	✓			
0.773	✓	✓		
0.767	✓		✓	
0.769		✓		
0.756			✓	
0.770		✓	✓	
0.776	✓	✓	✓	
0.773				✓
<b>0.798</b>	✓	✓	✓	✓

visualization of different sensory modalities when the robot *drops* a bottle is provided in Figure 3. Figure 3a illustrates the torques of 7 joints of the robot and 3 end-effector positions over time. Figure 3b shows the spectrogram of the auditory data. We use the Fast Fourier Transform to convert the raw signal into a representation in the frequency domain. Figure 3c shows the 3-axis accelerometer readings.

TABLE II: Modalities loss with and w/o auxiliary training, **aux** refers to auxiliary training and **no aux** refers to no auxiliary training.

	vision (SSIM)		haptic (MSE)	audio (MSE)	vibro (MSE)
	aux	no aux	aux		
vision	0.756		-	-	-
vision+haptic	0.785	0.764	0.282	-	-
vision+haptic+audio	0.796	0.791	0.246	0.042	-
vision+haptic+audio+vibro	0.798	0.795	0.244	0.041	0.739

### A. Training the Network with All Behaviors

The first experiment is to evaluate the framework in the all-behavior setting. Unlike the model in [2], which only uses one behavior (*push*), in the presented work, we trained the model on data spanning all 9 exploratory behaviors and evaluated it on novel unseen objects that were not seen during training. In this setting, we first show an illustrative example which describes the qualitative results of using multi-modal perceptions and a vision-only model compared with ground truth. Then we quantitatively evaluate the model performance with regards to different numbers of used sensory modalities. Furthermore, we study the model’s performance when the behavior type (*e.g. grasp vs. push*) is added as a categorical feature to the network. Note that except when explicitly indicated, the behavior category feature is used as input for the experiments.

**Illustrative Example.** Figure 4 shows the qualitative reconstruction performance of the proposed method and vision-only model [2] compared to ground-truth when the robot arm uses different behaviors (*push, lift*) to interact with objects. We observe that predicted frames using multi-modal are much less blurry. Furthermore, this figure demonstrates that the proposed method better captures the motion and can localize the object appearance with more precision especially in multiple steps into the future (*e.g. see location of robot arm and the object for push behavior, frame No. 16*).

**Quantitative Reconstruction Performance.** Figure 5a illustrates the performance of the network when integrated with different combinations of modalities compared to the vision-only method [2]. The results show that utilizing the network with multi-modal perceptions substantially increases the performance of the predicted frames. Note the gap between vision only and any combination of multi-modal escalates for further future frames. Meanwhile, as expected, the quality of prediction in all models decreases over time as errors accumulate. To avoid overfitting, we train the model with different channels in each layer and explore the effect of the model’s size on the performance. The baseline model explored in [2] contains 12.5M parameters, based on which we extend other modality sub-networks and reached 13.6M parameters. The number of associated parameters for the additional modalities are much less for two reasons. First the dimensions of other modalities are smaller compared to the vision. Second a deeper network is used for the visual branch.

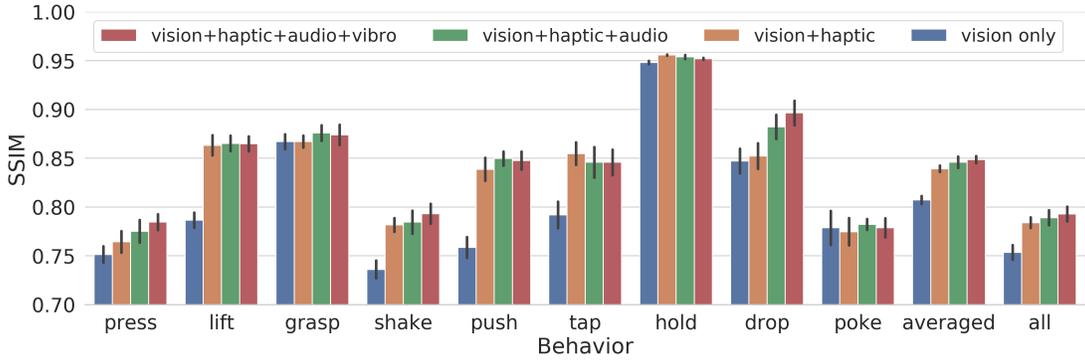


Fig. 6: Investigating the performance of different combinations of modalities per individual behavior

In another set of experiments, we investigate the effect of adding the behavior type as an input feature to the model. Figure 5b contrasts the model when it is trained with and without behavior. This figure shows the model performs better when the behavior is added as an input feature.

We demonstrate the contribution of each modality to the improvement of the model prediction via an ablation study. Table I shows the average SSIM over all time steps. The highest performance is obtained by integrating all modalities into the model. We also observe that in our dataset, haptic, audio, and behavior category share comparable contributions, while adding vibrotactile modality does not necessarily benefit the performance in this case, and sometimes it adds noise to the model which leads to performance degradation.

### B. Training Behavior-specific Models

We also investigate the performance of the model when trained and evaluated on an individual specific behavior. In this section, we ran the experiments with each behavior individually, yielding 9 models for each combination of modalities. We evaluate the performance of each model separately and also the averaged performance over all 9 behaviors. Furthermore, we compare the averaged performance to the model trained in section IV-A under the same combination setting. Finally, we explore how each behavior model performs differently from the others and investigate how they benefit from the additional modalities. Figure 6 shows the comparison between vision versus vision+haptic, vision+haptic+audio and vision+haptic+audio+vibrotactile for individual behaviors in terms of SSIM.

By comparing a different combination of modalities within each behavior, we observe that for 6 out of 9 behaviors, the model benefits from other modalities, especially, haptic. By contrasting the same modality setting across different behaviors, we notice that some behaviors (*lift*, *grasp*, *hold* and *drop*) pose an easier next-frame prediction challenge than others. We also observe that for tasks with discrete events (*e.g.* *drop*), the audio and tactile modalities are very helpful for predicting future frames; however, for contact behaviors, the haptic modality is significantly more helpful than audio

and tactile feedback. Furthermore, by integrating 9 separate models, we evaluate the averaged performance of the model (the 'averaged' column in figure 6). The averaged performance of the behavior-specific models is higher than that of the model trained simultaneously on all behaviors as described in Section IV-A, shown in the rightmost column.

### C. Predicting Future Frames of Auxiliary Modalities

Another novelty of the proposed framework is predicting future frames of modalities other than vision. Predicting other modalities can sometimes be useful (*e.g.* comparing the difference between predicted audio and the observed audio modality to identify abnormal events as they happen). In this subsection, we investigate the performance of these auxiliary tasks and whether learning them helps improve visual next-frame prediction. We evaluate vision modality prediction in two settings: **with auxiliary training** and **without auxiliary training** settings and assess the performance of the next-frame prediction model for the non-visual modalities under the **with auxiliary training** setting.

Table II shows that auxiliary training of haptic modality enhances vision prediction while auxiliary training of audio and vibrotactile modalities does not necessarily contribute to improve visual next-frame prediction. Furthermore, this table shows that the audio modality contributes to the prediction of the haptic, while the vibrotactile modality seems to have little influence on predicting haptic and audio modalities.

## V. CONCLUSION AND FUTURE WORK

In this work, we developed a predictive framework incorporating multiple sensory modalities to help solve the next-frame prediction problem. Our experiments show that utilizing the network architecture with additional haptic, auditory, and tactile inputs achieves the best results compared to a state-of-the-art vision-only baseline. Furthermore, in this paper, we proposed the use of auxiliary tasks (predicting future haptic, audio, and vibrotactile signals) and showed that learning such tasks also improves visual next-frame prediction.

One limitation of our framework is that it is trained on only one robot. Since different robots have different morphologies and different sensor suites, the learned knowledge cannot be

directly used by another robot. An interesting avenue for future work is to extend transfer learning methodologies (e.g., [38], [37]) as to enable a robot to bootstrap its sensorimotor learning process with knowledge learned by another robot. Another viable direction for future work is to integrate the multisensory next-frame prediction methodology described here with reinforcement learning methods for object manipulation tasks.

## REFERENCES

- [1] M. V. B. O. Sigaud and G. P. G. Baldassarre, *Anticipatory behavior in adaptive learning systems*. Springer, 2007.
- [2] C. Finn, I. Goodfellow, and S. Levine, “Unsupervised learning for physical interaction through video prediction,” in *Advances in neural information processing systems*, 2016, pp. 64–72.
- [3] T. Wilcox, R. Woods, C. Chapa, and S. McCurry, “Multisensory exploration and object individuation in infancy,” *Developmental Psychology*, vol. 43, no. 2, p. 479, 2007.
- [4] M. O. Ernst and H. H. Bühlhoff, “Merging the senses into a robust percept,” *Trends in cognitive sciences*, vol. 8, no. 4, pp. 162–169, 2004.
- [5] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. S. Sukhatme, “Interactive perception: Leveraging action in perception and perception in action,” *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1273–1291, 2017.
- [6] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, “A review of tactile information: Perception and action through touch,” *IEEE Transactions on Robotics*, 2020.
- [7] G. Tatiya and J. Sinapov, “Deep multi-sensory object category recognition using interactive behavioral exploration,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7872–7878.
- [8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *ICML*, 2011.
- [9] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, “Deep learning for tactile understanding from visual and haptic data,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 536–543.
- [10] K. Zhang, M. Sharma, M. Veloso, and O. Kroemer, “Leveraging multimodal haptic sensory data for robust cutting,” in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2019, pp. 409–416.
- [11] F. Pastor, J. García-González, J. M. Gandarias, D. Medina, P. Closas, A. J. García-Cerezo, and J. M. Gómez-de Gabriel, “Bayesian and neural inference on lstm-based object recognition from tactile and kinesthetic information,” *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 231–238, 2020.
- [12] S. Chitta, J. Sturm, M. Piccoli, and W. Burgard, “Tactile sensing for mobile manipulation,” *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 558–568, 2011.
- [13] Y. Zhang, W. Yuan, Z. Kan, and M. Y. Wang, “Towards learning to detect and predict contact events on vision-based tactile sensors,” *arXiv preprint arXiv:1910.03973*, 2019.
- [14] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith, and A. Stoytchev, “Interactive object recognition using proprioceptive and auditory feedback,” *The International Journal of Robotics Research*, vol. 30, no. 10, pp. 1250–1262, 2011.
- [15] S. Jin, H. Liu, B. Wang, and F. Sun, “Open-environment robotic acoustic perception for object recognition,” *Frontiers in Neurorobotics*, vol. 13, p. 96, 2019.
- [16] D. Gandhi, A. Gupta, and L. Pinto, “Swoosh! Rattle! Thump! - Actions that Sound,” in *Proceedings of Robotics: Science and Systems*, Corvallis, Oregon, USA, July 2020.
- [17] R. Braud, A. Giagkos, P. Shaw, M. Lee, and Q. Shen, “Robot multimodal object perception and recognition: Synthetic maturation of sensorimotor learning in embodied systems,” *IEEE Transactions on Cognitive and Developmental Systems*, 2020.
- [18] G. Tatiya, R. Hosseini, M. C. Hughes, and J. Sinapov, “Sensorimotor cross-behavior knowledge transfer for grounded category recognition,” in *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2019, pp. 1–6.
- [19] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker, “Robotic learning of haptic adjectives through physical interaction,” *Robotics and Autonomous Systems*, vol. 63, pp. 279–292, 2015.
- [20] S. Amiri, S. Wei, S. Zhang, J. Sinapov, J. Thomason, and P. Stone, “Multi-modal predicate identification using dynamically learned robot controllers,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.
- [21] B. Richardson and K. Kuchenbecker, “Improving haptic adjective recognition with unsupervised feature learning,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- [22] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidson, J. Hart, P. Stone, and R. Mooney, “Jointly improving parsing and perception for natural language commands through human-robot dialog,” *Journal of Artificial Intelligence Research*, vol. 67, pp. 327–374, 2020.
- [23] J. Yuen and A. Torralba, “A data-driven approach for event prediction,” in *European Conference on Computer Vision*. Springer, 2010, pp. 707–720.
- [24] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, “Video (language) modeling: a baseline for generative models of natural videos,” *arXiv preprint arXiv:1412.6604*, 2014.
- [25] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using lstms,” in *International conference on machine learning*, 2015, pp. 843–852.
- [26] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” *arXiv preprint arXiv:1511.05440*, 2015.
- [27] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, “Action-conditional video prediction using deep networks in atari games,” in *Advances in neural information processing systems*, 2015, pp. 2863–2871.
- [28] X. Liang, L. Lee, W. Dai, and E. P. Xing, “Dual motion gan for future-flow embedded video prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1744–1752.
- [29] N. Wichers, R. Villegas, D. Erhan, and H. Lee, “Hierarchical long-term video prediction without supervision,” *arXiv preprint arXiv:1806.04768*, 2018.
- [30] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, “Stochastic adversarial video prediction,” *arXiv preprint arXiv:1804.01523*, 2018.
- [31] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, “Stochastic variational video prediction,” *arXiv preprint arXiv:1710.11252*, 2017.
- [32] W. Lotter, G. Kreiman, and D. Cox, “Deep predictive coding networks for video prediction and unsupervised learning,” *arXiv preprint arXiv:1605.08104*, 2016.
- [33] M. Jaderberg, K. Simonyan, A. Zisserman, et al., “Spatial transformer networks,” in *Advances in neural information processing systems*, 2015, pp. 2017–2025.
- [34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.
- [35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [36] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.
- [37] G. Tatiya, R. Hosseini, M. C. Hughes, and J. Sinapov, “A framework for sensorimotor cross-perception and cross-behavior knowledge transfer for object categorization,” *Frontiers in Robotics and AI*, vol. 7, p. 137, 2020.
- [38] G. Tatiya, Y. Shukla, M. Edegbare, and J. Sinapov, “Haptic knowledge transfer between heterogeneous robots using kernel manifold alignment,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.