# MUTEX: Learning Unified Policies from Multimodal Task Specifications

**Anonymous Author(s)**
Affiliation
Address
email

**Abstract:** Humans communicate the goal of a task to their team partners using different modalities: speech, text, images... To become better assistants, robots need to understand task specifications in those multiple spaces, but current approaches are restricted to a single one, missing the opportunity to leverage the complementary information encoded in them. We present MUTEX, a novel approach to learning unified policies from multimodal task specifications that leverages cross-modal information thanks to its two-stage training procedure combining masked modeling and cross-modal matching, and its novel architecture based on transformers. After training, MUTEX can follow a task specification in any of the six learned modalities (video demonstrations, goal images, text goal descriptions, text instructions, speech goal descriptions, and speech instructions), or combinations of specifications in several of them. We train and evaluate empirically the benefits of MUTEX in a novel dataset with 100 tasks in simulation and 40 tasks in the real world, annotated with multiple instances of task specifications in the six modalities, and observe improved performance over methods trained specifically for every single modality. More information at https://sites.google.com/view/mutex

**Keywords:** Multimodal Learning, Task Specification, Manipulation

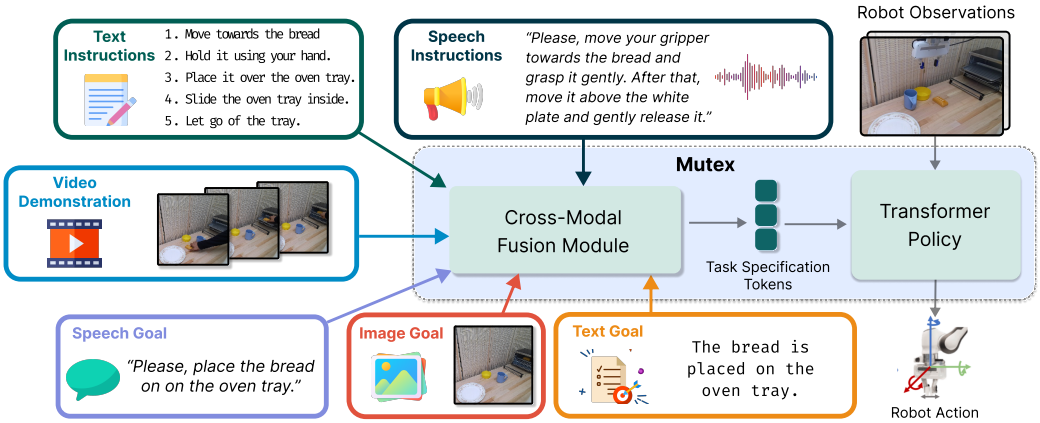


Figure 1: **Overview.** We introduce MUTEX, a unified policy that learns to perform tasks based on task specifications in multiple modalities: text, image, video, and speech, including instructions and goal descriptions. Based on a novel two-stage cross-modal representation learning procedure, MUTEX learns to leverage information across modalities to become more capable at executing tasks specified by any single modality than methods trained specifically for each one.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

# 1 Introduction

When working in a team, humans make use of different modalities to specify tasks and improve communications, *e.g.*, sharing high-level task goals ("Let's cook a meal!"), verbal instructions ("We will go to the kitchen, get the pot from the cabinet, and then put it on the stove."), or fine-grained visual demonstrations (showing a cooking video). Human-robot teams should aspire to a similar level of understanding. While recent research in robot learning has studied various modalities for specifying robotic tasks, including text, images, speech, and videos, most previous studies have treated these individual modalities as separate problems, such as language-conditioned policy learning [1, 2, 3, 4], instruction following [5], visual goal-reaching [6, 7, 8, 9], and imitation from video demonstrations [10, 11, 12]. Consequently, these approaches lead to siloed systems tailored to individual task specification modalities.

A burgeoning body of interdisciplinary AI research has suggested that joint learning across multiple modalities, such as image and text [13, 14] and vision and touch [15], gives rise to richer and more effective representations that improve understanding of individual modalities. These results align with findings in cognitive science and psychology, which suggest that incorporating multimodal cues (*e.g.*, visual and verbal) into human learning processes enhances learning quality over using individual modalities alone [16, 17]. Drawing inspiration from the effectiveness of cross-modal learning in prior work, our goal is to develop **unified policies capable of reasoning about multi-modal task specifications for diverse manipulation tasks**, where each task will be defined in a single modality that changes from task to task. We seek to harness the complementary strengths of different modalities — some providing compact and high-level information like goal descriptions, while others providing fine-grained information like step-by-step video demonstrations — thereby improving the model's ability to execute tasks from task specifications of different modalities.

A key challenge associated with learning across modalities is effectively leveraging cross-modal information to reinforce each other. Prior work either focuses on enhancing language understanding using visual data [18, 19] or requires both language and video demonstrations to specify tasks [20]. These approaches typically cover one or two modalities, failing to encompass the diverse ways in which humans express their goals and intent. Furthermore, these methods lack an explicit learning mechanism to incentivize cross-modal interaction, thereby failing to enrich the representation of each modality.

To effectively learn from different task specification modalities, we improve upon two representation learning techniques, **masked modeling** [13, 21, 22] and **cross-modal matching** [23, 24], to foster cross-modal interaction through a shared embedding space. Firstly, we exploit the complementary strengths of each modality, where language and speech specifications help in focusing important features from visual specification while visual specification helps to ground the language and speech, through cross-modal interactions. This is accomplished by masking certain parts of the input task specification and predicting the missing elements with the help of others. This loss is optimized in tandem with the policy learning objective (*i.e.*, behavior cloning) so that the representation of the task specification also captures action-relevant information. After learning a richer, more-informed representation of each modality, we bring representations of different modalities to a common space [13, 23, 21, 22]. Unlike prior work that maps visual specifications to language embeddings [25], we exploit the fact that human video demonstrations contain more fine-grained information about the task. We enrich the representations of other modalities by matching them with the richer video representation. This cross-modal matching gives rise to multimodal representations that are compact yet informative, which can be used to execute tasks specified by any modality.

For the model to execute a task specified by any modality, it must be able to handle the variable input length of task specification tokens. Meanwhile, it must be able to predict robot actions alongside a variable number of masked signals for each modality. To achieve this, we use a Perceiver-style encoder [26] where a variable number of task specification tokens are attended with fixed history of robot observations using the cross-attention mechanisms. The embeddings obtained from the

encoder, are then passed through the Perceiver-style decoder [27] to predict robot actions and masked signals for individual modalities. This architecture design allows the model to learn a policy that can execute tasks specified by any single modality or arbitrary aggregation of several of them.

In summary, we introduce MUTEX (**MU**ltimodal **T**ask specification for robot **EX**ecution), a unified policy capable of executing tasks specified as goal descriptions or more detailed instructions in natural language, speech, or visual modalities. MUTEX is not only more versatile thanks to its multi-modal understanding, but it is also more performant, improving the robustness of the task execution on every single modality. This is demonstrated empirically with a comprehensive evaluation benchmark including diverse manipulation tasks, 100 in simulation and over 40 in real-world, leading to 6000 evaluated trials in simulation per method and over 50 trials evaluated per method in real-world, totaling 600 evaluation trials. Real-world evaluation indicates that MUTEX can effectively interpret human video demonstrations and perform tasks successfully with the robot, albeit with morphology differences. As part of this effort, we provide a large real-world dataset of 40 tasks with 30 trajectories for each task containing tasks like "putting bread on a microwave tray and closing it", "opening an air fryer and putting hot dogs in it", or "placing the book in the front compartment of the caddy", specified with multiple rich multimodal task specifications: text goal description, text instruction, goal image, video specification, speech goal description, and speech instruction. We hope this new dataset will support further research in multimodal task specification.

## 2 Related Work

MUTEX learns to interpret task specifications in multiple modalities for a diverse set of demonstrated tasks through a cross-modal representation learning procedure. In this section, we review different forms of task specification commonly seen in the robot manipulation literature.

The simplest way to specify tasks is through one-hot encoding vectors [28]. However, this representation is restricted to a predefined set of tasks that cannot generalize to new ones, ultimately resulting in suboptimal policies [25]. In contrast, methods that consume tasks specified in natural language [1, 25, 5, 3, 4, 29, 30, 31] have shown improved multi-task generalization thanks to a richer semantic task embedding. However, learning with natural language specification can be ineffective as it requires grounding in the observation and action space of the robot [32, 33] and the symbolic nature of language poses a challenge for tasks that require more granular or accurate descriptions [34]. These are not issues for methods that consume visual specification of tasks (images [7, 9, 8] or videos [10, 11, 25]), but these methods may suffer due to irrelevant information present in the specification (a simple task of moving an object also specifies the location of other objects in the scene [2, 35]), and, due to the higher dimensionality of images, these methods tend to require larger amounts of data and demonstrations.

Closer to ours are methods that use multi-modal task specifications. Some of them leverage multi-modal data during training but lead to final policies that can only consume one single type of modality [18, 19, 25]. Others demonstrate that policies consuming multi-modal specifications can generalize to new tasks in a 1-shot fashion but require each new task to be specified in all modalities [20]. A recent group of methods consumes single task specifications defined in a combination of modalities (part text, part image, ...) achieving more flexibility and potentially alleviating grounding problems [34]. None of the prior works offer the flexibility to specify the task using any individual modality, nor support as many different modalities as MUTEX (visual, language, speech), missing the opportunity to leverage their complementary information through cross-modal learning.

## 3 MUTEX Model

Our goal is to learn a unified policy that performs a diverse set of tasks (and generalizes to new ones) based on a dataset of demonstrations annotated with multi-modal task specifications including language, speech, and visual specification. We assume that, during test time, the task to perform will be specified in a single modality that can change from task to task. Formally, for each task,

$T_i \in \{T_1, T_2, \ldots, T_n\}$, we assume that the agent learns from a set of expert demonstrations obtained via human teleoperation, $D_i = \{d_i^1, d_i^2, \ldots, d_i^m\}$, where $m$ is the number of demonstrations for task $T_i$, forming an entire dataset of demonstrations, $D$. Each demonstration $d_i^j$ present the form of a sequence of observations and expert actions, $d_i^j = [(o_1, a_1)_i^j, \ldots (o_T, a_T)_i^j]$.

The goal of each training task $T_i$ is specified by $k$ alternative task descriptions, $t$, in six different forms within three modalities: text instructions, $t \in \{L_i^1, L_i^2, \ldots, L_i^k\}$, text goal description, $t \in \{l_i^1, l_i^2, \ldots, l_i^k\}$, video demonstration, $t \in \{V_i^1, V_i^2, \ldots, V_i^k\}$, final goal image, $t \in \{v_i^1, v_i^2, \ldots, v_i^k\}$, speech instructions, $t \in \{S_i^1, S_i^2, \ldots, S_i^k\}$, and speech goal description, $t \in \{s_i^1, s_i^2, \ldots, s_i^k\}$. Note that we characterize the modalities with the letters $L/l$, $V/v$, and $S/s$, and make use of capital letters to denote *path specifications* and small letters to denote *goal state specifications*. Our goal is to learn a unified policy, $\pi(o|t) = a$, that outputs continuous actions, $a \in A$, given current observations, $o_t \in O$, conditioned on a task specification in one of the possible modalities, $t \in L|l|V|v|S|s$. We aim to create a policy that not only performs the $n$ tasks in the training dataset $D$ (seen tasks), even in new initial conditions (*e.g.* positions of the objects) but also generalizes to novel tasks specified by previously unseen task descriptions.

### 3.1 MUTEX Training Procedure

Our goal is not only to obtain a unified policy that understands task specifications in different modalities, but we would also like to improve the policy performance when a task is specified on every single modality by exploiting cross-modal interactions during training. To that end, we leverage two representation learning techniques that we integrate sequentially in MUTEX's training procedure: **1) Mask Modeling** to promote cross-modal interactions of all modalities into a shared learned latent space, and **2) Cross-Modal Matching** to enrich each modality with information of the information-denser one. Both stages of our procedure are combined with a behavior cloning objective for policy learning to ensure that the learned representation contains relevant information for the agent to perform the manipulation task. The overview of the proposed approach is shown in Fig. 2. In the following sections, we describe each element in detail.

#### 3.1.1 Masked Modeling for Cross-Modal Learning

In the first stage of MUTEX's training procedure, we would like to promote cross-modal interactions between the model components interpreting the different modalities: text instructions ($L$), text goal ($l$), video demonstration ($V$), image goal ($v$), speech instructions ($S$), and speech goal ($s$). To that end, and inspired by the success in other cross-modal learning tasks [38, 14, 39, 40], we mask certain tokens or features of each modality and learn to predict them with the help of other modalities. This enforces the model to use the information from other modalities to enrich the representation of each one of them. Additionally, during testing we use a task specification in only one or a subset of the modalities; therefore, we recreate these conditions in our training procedure to obtain robust single-modality representations.

Specifically, at each iteration of the training process, we randomly select task specifications in one or a subset of modalities. If more than one modality is selected, we mask certain parts of each modality. We request from MUTEX to predict the masked parts alongside the action values the expert demonstrated at each step. Assuming $y_1, y_2, \ldots, y_k$ are the masked tokens in one modality, the loss we train with is given by:

$$\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}(y_i, \hat{y}_i) \tag{1}$$

Different modalities would mask at different stages of the model (input or features) and use a different metric loss to measure prediction error ($\ell_1$ or $\ell_2$).

For **masked language modeling** (masking elements of a language goal description or language instructions), we mask out entire words [41] that are then passed through pretrained CLIP language model [36] to extract the masked features to predict. We use the standard cross-entropy loss between

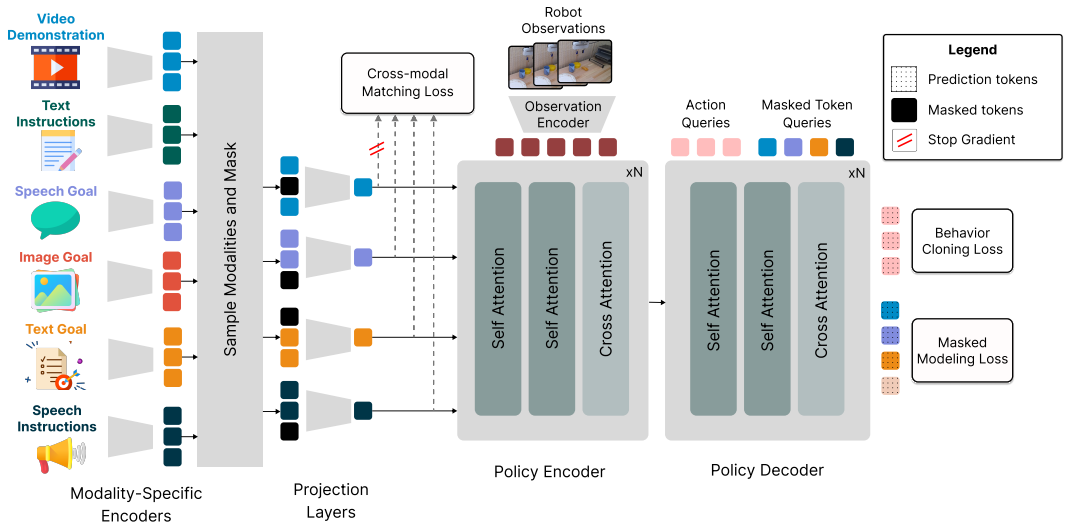


Figure 2: **MUTEX's Model Architecture and Training Losses**. Task specifications in each modality are encoded with pretrained modality-specific encoders, such as OpenAI's CLIP and Whisper models [36, 37]. During the first stage of training, one or more of these modalities are randomly selected and masked before being passed to projection layers and input to MUTEX's transformer encoder. The resultant tokens are combined with observation tokens through $N$ blocks of self- and cross-attention layers. The hidden state of the encoder is passed to MUTEX's transformer encoder that is queried for actions (behavior learning loss) and the masked features and tokens (masked modeling loss), promoting action-specific cross-modal learning. In the second stage of training, all modalities are enriched with information from video features through a cross-modal matching loss. At test time, observations and a task specification of one modality are provided, based on which MUTEX predicts closed-loop actions to achieve the task.

the predicted and ground truth tokens, *i.e.*, $\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$, where $N$ is the number of tokens in the vocabulary. For **masked visual modeling** (masking elements of an image goal or a video instruction), we mask out intermediate regions of the features obtained from a pre-trained CLIP model [36] and, following prior works in vision-language modeling [42], we employ a simple $\ell_1$-regression loss between the predicted and ground truth features, $y$ and $\hat{y}$. Finally, for **masked speech modeling** (masking elements of a speech goal description or set of instructions), we use a similar approach to visual modeling, but using features from a pre-trained Whisper model [37], instead. We also use an $\ell_1$-regression loss for the predictions.

#### 3.1.2 Cross-Modal Matching for Richer Representations

In the second stage of MUTEX's training procedure, we would like to enrich the common embedding space for each task modality by associating it with the features of the information-richer one. To that end, in contrast to prior works that use a cross-modal contrastive loss to learn a common embedding space [23, 43], we use a simple $\ell_2$ loss to pull the representations of all modalities towards the one with more information and better performance. Video specifications contain the most information, leading to more elucidative and stronger features; therefore, we enrich other modalities with information from the video representation space obtained after cross-modal learning.

Concretely, let $f_L$, $f_l$, $f_v$, $f_s$, $f_S$ be the feature representations of the language instruction, language goal, image goal, speech goal, and speech instructions, and $f_V$ the feature representation of the video demonstration for the same task. Our cross-modal matching loss is given by:

$$\mathcal{L}_{match} = \mathcal{L}_{\ell_2}(f_L, f_V) + \mathcal{L}_{\ell_2}(f_l, f_V) + \mathcal{L}_{\ell_2}(f_S, f_V) + \mathcal{L}_{\ell_2}(f_s, f_V) + \mathcal{L}_{\ell_2}(f_s, f_V) \tag{2}$$

where $\mathcal{L}_{\ell_2}$ is the $\ell_2$-regression loss. Gradients from this loss are not back propagated to the part of MUTEX that encodes the video features, leaving them unchanged.

### 3.2 MUTEX Architecture

The training process delineated above requires a suitable model architecture capable of propagating and encoding the cross-modal information from multiple modalities. MUTEX's model consists of three main components (see Fig. 2): 1) Modality Specific Encoders that map input modalities to task-specific tokens, 2) a Policy Encoder that takes in the task-specific tokens and robot observations and outputs hidden states, and 3) a Policy Decoder that takes in the hidden states along with decoder queries, and outputs corresponding features corresponding to the queries.

MUTEX's **modality-specific encoders** extract tokens from input task specification using a fixed, pre-trained large model which helps us to extract semantically meaningful representation from the input modality. In order to learn representations that are grounded to the robot observation and action space, these features are passed through a projection layer consisting of a simple MLP or single attention layer before passing it to the policy encoder. MUTEX's **policy encoder** aims to fuse information obtained from multiple task specification modalities as well as robot observations. To successfully fuse the information we adopt a transformer-based architecture consisting of stacked cross- and self-attention layers. In the cross-attention layers, the queries are generated from robot observations whereas keys and values are generated from task specification tokens [34, 26]. The output of the encoder is passed on to the policy decoder. The output from the policy encoder is enriched with information obtained from different task specification modalities. However, MUTEX requires the policy to output features for predicting action values as well as a variable number of masked tokens. This motivates us to adopt a Perceiver Decoder [27] architecture as MUTEX's **policy decoder** to leverage learnable queries and output only the information corresponding to input queries. Thereby, the decoder features corresponding to action prediction are passed through an MLP to predict continuous action values that estimate a Gaussian Mixture Model with 5 modes. The decoder features corresponding to masked token queries are passed through separate MLPs to predict token values or features.

## 4 Multimodal Task Specification Dataset

As part of our efforts to develop a unified multi-task imitation learning policy from multimodal task specifications, we collected a novel dataset of tasks with multiple definitions per modality. Our dataset follows the description provided in Sec. 3: $n$ tasks, with $m$ demonstrations per task and $k$ different task definitions per task in each of the six formats: text instructions and text goal description, video demonstration, image goal, speech instruction, and speech goal description. Our dataset is comprised of a set of simulated tasks and a set of real-world tasks (see Fig. 3).

In simulation, we extended the LIBERO-100 [44] consisting of $n = 100$ tasks each with $m = 50$ expert trajectories provided by original work with $k = 10$ task specifications per modality. Due to the inevitable sim2real gap, video demonstrations in simulation are collected by teleoperating the simulated robot instead of directly by a human performing the tasks.

In real-world, we collected a novel dataset with $n = 40$ tasks ranging from pick and place tasks such as "put the bread on the white plate", "pick up the bowl at the back of the scene and place it inside the top drawer" to more contact-rich tasks such as "open the air fryer and put the bowl with dogs in it", "take out the tray of the oven, and put the bread on it." All tasks are collected in the same environment but involve different objects from a set of 18 objects. We use a Franka Panda robot and an Intel RealSense camera for an eye-in-hand view whereas a Kinect Azure workspace camera. Each task is demonstrated with $m = 30$ human-collected trajectories using a 3D spacemouse teleoperation device. Each task is annotated with $k = 10$ different task specifications for each modality. Video demonstrations specifying the tasks are performed directly by a human.

To generate ten diverse text goal descriptions and instructions, we make use of ChatGPT with prompts to generate alternative descriptions that we manually filter to avoid using synonyms that do not match the right task-relevant objects (*e.g.*, using *plate* as a synonym for *bowl* when there are other plates in the scene). We request audio signals from several characters from the Amazon Polly

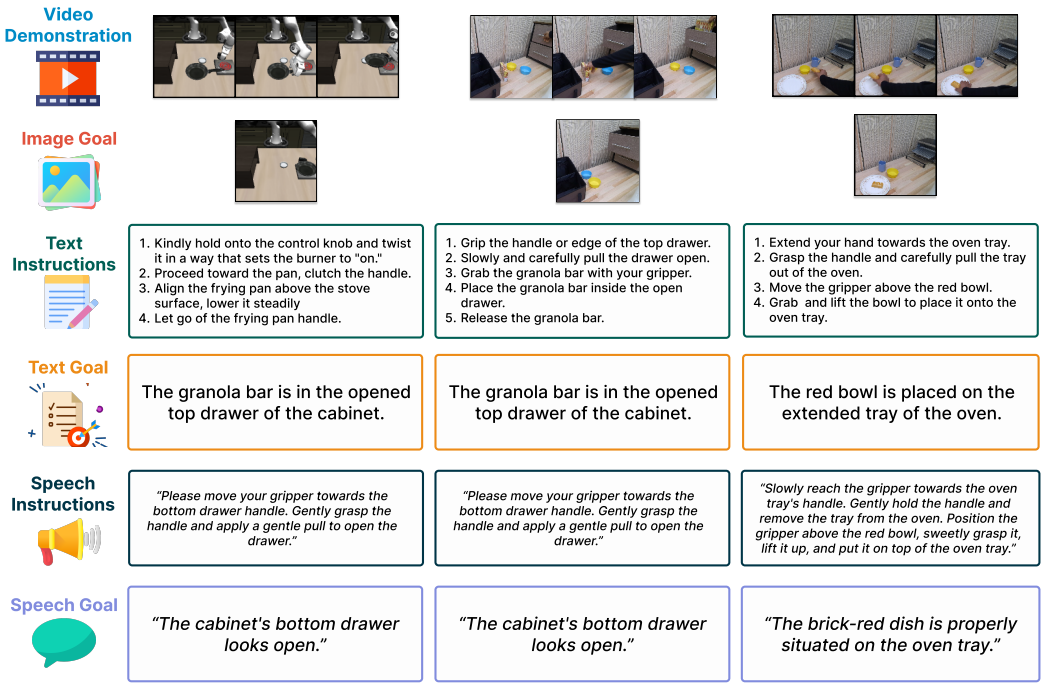


Figure 3: **MUTEX Multimodal Task Specification Dataset.** We provide a dataset comprising **100** simulated tasks (example in the first column) based on LIBERO-100 [44] and **40** real-world tasks (examples in the second and third columns), annotated with **50** and **30** demonstrations per task, respectively. We annotate each task with **ten** alternative task specifications in each of the **six** following modalities (rows from top to bottom): video demonstration, image goal, text instructions, text goal, speech instructions, and speech goal.

service to generate ten diverse speech descriptions. Empirically, we find that sufficient variability in the audio signal is critical to producing robust solutions.

Our dataset of multimodal task specifications will be made public upon acceptance.

## 5 Experimental Evaluation

**Experimental Setup:** We use our developed dataset of multimodal task specifications for the evaluation, with a split of 80%/20% tasks for training and testing. For each tested task, we evaluate 20 trials in simulation and 10 trials in the real world. We then average results for all unseen tasks.

In our evaluation, we compare MUTEX to models trained specifically for each modality using only task specifications in that modality. This resembles most existing prior works in video-based imitation learning [10, 11, 12], and goal-conditioned imitation learning [7, 45]. When training MUTEX and the baselines, we use the same architecture design and hyperparameters for fair comparison.

**Experiments:** In our experimental evaluation, we aim to answer the following questions: *1) Does a unified policy capable of executing tasks across multiple modalities outperform methods trained specifically from and for each individual modality?* Table 1 and Table 2 summarize the results of our evaluation in simulation and the real world, respectively. In both cases, we observe a significant improvement from using our unified policy MUTEX compared to modality-specific models indicating that the cross-modal representation learning procedure from MUTEX is able to leverage more information from other modalities.

*2) What is the relative importance of the two stages of the* MUTEX *training procedure? Is it important to perform the stages consecutively?* Table 1 includes a comparison of the results obtained with

the two-staged training procedure consecutively (MUTEX) compared to training with both stages simultaneously (joint training), training without the first stage (no masked modeling) or without the second stage (no cross-modal matching). We observe that the largest drop in performance comes from training without masked modeling, indicating that this step is critical to learn cross-modal information from different task specifications. The cross-modal matching provides an additional boost except for video demonstrations (we match to that modality) and for speech instructions, where we observe a small drop.

*3) Does the performance increase significantly when tasks are specified with multiple modalities?* One of the advantages of MUTEX is that it can execute tasks with a single specification in any of the learned modalities, or with multiple specifications in several of the modalities. We evaluated using combinations of *Text Goal* + *Speech Goal*, *Text Goal* + *Image Goal* and *Speech Instructions* + *Video Demonstration* and obtain, $50.1$, $59.2$, and $59.6$ success rates, respectively. These values are close to the performance using one single specification in the best of the two modalities, indicating that the additional modality is not providing any extra information. We hypothesize this is because all possible cross-modal information has already been learned by MUTEX. Interestingly, when using specifications in *all modalities*, the success rate is $60.1$, lower than when using only *Image goal* or *Video Demonstration*. This is possibly due to the increased complexity of interpreting multiple task specifications.

| Method | Text Goal | Text Instructions | Image Goal | Video Demonstration | Speech Goal | Speech Instructions |
|---|---|---|---|---|---|---|
| Modality-Specific | $41.7 \pm 8.0$ | $39.9 \pm 2.8$ | $58.7 \pm 4.9$ | $62.0 \pm 5.2$ | $22.3 \pm 1.2$ | $28.4 \pm 5.5$ |
| MUTEX (joint training) | 45.2 | 42.9 | 50.4 | 53.5 | 32.8 | 35.1 |
| MUTEX (no masked modeling) | $27.4 \pm 0.7$ | $30.9 \pm 0.2$ | $36.3 \pm 0.1$ | $38.5 \pm 2.2$ | $21.7 \pm 2.6$ | $23.0 \pm 0.7$ |
| MUTEX (no cross-modal matching) | $43.5 \pm 8.9$ | $39.4 \pm 2.2$ | $60.1 \pm 6.4$ | $\mathbf{63.2 \pm 6.3}$ | $36.7 \pm 4.3$ | $\mathbf{46.8 \pm 5.5}$ |
| MUTEX | $\mathbf{50.1 \pm 7.8}$ | $\mathbf{53.0 \pm 2.2}$ | $\mathbf{61.6 \pm 6.4}$ | $\mathbf{63.2 \pm 6.3}$ | $\mathbf{40.9 \pm 8.1}$ | $46.0 \pm 5.2$ |

Table 1: Success Rate in the Multimodal Task Specification Dataset in Simulation.

| Method | Text Goal | Text Instructions | Image Goal | Video Demonstration | Speech Goal | Speech Instructions |
|---|---|---|---|---|---|---|
| Modality-Specific | 48 | 48 | 50 | 52 | 30 | 44 |
| MUTEX | **58** | **58** | **58** | **56** | **36** | **54** |

Table 2: Success Rate in the Multimodal Task Specification Dataset in Real-World.

## 6 Conclusion, Limitations, and Future Work

We proposed MUTEX, a general model to learn unified policies for multimodal task specifications in multi-task robot imitation learning for manipulation that can execute tasks specified using any of the learned modalities by using cross-modal representation learning in a two-stage procedure. We demonstrate with multiple experiments in simulation and real-world that an agent trained with multiple modalities, each of which reinforces the others, produces a more robust and versatile policy in each modality. However, MUTEX is not without limitations. First, we assume paired access to all the task specification modalities, which may be hard to obtain for new tasks. Second, for speech, we used clean signals generated by Amazon Polly. Real-world speech can be noisier and more difficult to understand. And third, our video and image goals are provided from the same workspace as the task to be executed. Having a policy that can execute task specified "in the wild" visual goal or demonstration will create additional challenges. We plan to explore how to augment generalization by allowing different environments, which could open the door to the use of larger datasets of human tasks [46].

# References


[1] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

[2] S. Nair, E. Mitchell, K. Chen, S. Savarese, C. Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pages 1303–1315. PMLR, 2022.

[3] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[4] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[5] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. *arXiv preprint arXiv:2210.06407*, 2022.

[6] D. Pathak, P. Mahmoudieh, G. Luo, P. Agrawal, D. Chen, Y. Shentu, E. Shelhamer, J. Malik, A. A. Efros, and T. Darrell. Zero-shot visual imitation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2050–2053, 2018.

[7] S. Nasiriany, V. Pong, S. Lin, and S. Levine. Planning with goal-conditioned policies. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] Y. Cui, S. Niekum, A. Gupta, V. Kumar, and A. Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Learning for Dynamics and Control Conference*, pages 893–905. PMLR, 2022.

[9] Z. J. Cui, Y. Wang, N. Muhammad, L. Pinto, et al. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint arXiv:2210.10047*, 2022.

[10] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, and S. Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018.

[11] A. Bonardi, S. James, and A. J. Davison. Learning one-shot imitation from humans without humans. *IEEE Robotics and Automation Letters*, 5(2):3533–3539, 2020.

[12] S. James, M. Bloesch, and A. J. Davison. Task-embedded control networks for few-shot imitation learning. In *Conference on robot learning*, pages 783–795. PMLR, 2018.

[13] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020.

[14] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[15] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks, 2019.

[16] J. M. Clark and A. Paivio. Dual coding theory and education. *Educational psychology review*, 3:149–210, 1991.

[17] L. E. Bahrick and R. Lickliter. Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, 36(2):190, 2000.

[18] C. Lynch and P. Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[19] O. Mees, L. Hermann, and W. Burgard. What matters in language conditioned robotic imitation learning. *arXiv preprint arXiv:2204.06252*, 2022.

[20] A. Yu and R. J. Mooney. Using both demonstrations and language instructions to efficiently learn robotic tasks. *arXiv preprint arXiv:2210.04476*, 2022.

[21] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[22] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*, 2022.

[23] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[24] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[25] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[26] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General perception with iterative attention, 2021.

[27] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[28] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3758–3765. IEEE, 2018.

[29] H. Liu, L. Lee, K. Lee, and P. Abbeel. Instruction-following agents with jointly pre-trained vision-language models. *arXiv preprint arXiv:2210.13431*, 2022.

[30] P.-L. Guhur, S. Chen, R. G. Pinel, M. Tapaswi, I. Laptev, and C. Schmid. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pages 175–187. PMLR, 2023.

[31] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[32] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[33] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

[34] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

[35] P. Agrawal. The task specification problem. In *Conference on Robot Learning*, pages 1745–1751. PMLR, 2022.

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

[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[38] H. Tan and M. Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*, 2020.

[39] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang. Florence: A new foundation model for computer vision, 2021.

[40] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all, 2023.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[42] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[43] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[44] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023.

[45] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training, 2023.

[46] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.