Robo-MUTUAL: Robotic Multimodal Task Specification via Unimodal Learning

Jianxiong Li[1*], **Zhihao Wang**[2,1*], **Jinliang Zheng**[1,3*], **Xiaoai Zhou**[4,1], **Guanming Wang**[5,1], **Guanglu Song**[3], **Yu Liu**[3], **Jingjing Liu**[1], **Ya-Qin Zhang**[1], **Junzhi Yu**[2†] and **Xianyuan Zhan**[1,6†]
[1] AIR, Tsinghua University, [2] Peking University, [3] Sensetime Research
[4] University of Toronto, [5] UCL, [6] Shanghai AI Lab
li-jx21@mails.tsinghua.edu.cn
zhanxianyuan@mail.tsinghua.edu.cn

Abstract

Multimodal task specification is essential for enhanced robotic performance, where *Cross-modality Alignment* enables the robot to holistically understand complex task instructions. Directly annotating multimodal instructions for model training proves impractical, due to the sparsity of paired multimodal data. In this study, we demonstrate that by leveraging unimodal instructions abundant in real data, we can effectively teach robots to learn multimodal task specifications. First, we endow the robot with strong *Cross-modality Alignment* capabilities, by pretraining a robotic multimodal encoder using extensive out-of-domain data. Then, we employ two Collapse and Corrupt operations to further bridge the remaining modality gap in the learned multimodal representation. This approach projects different modalities of identical task goal as interchangeable representations, thus enabling accurate robotic operations within a well-aligned multimodal latent space. Evaluation across more than 130 tasks and 4000 evaluations on both simulated LIBERO benchmark and real robot platforms showcases the superior capabilities of our proposed framework, demonstrating significant advantage in overcoming data constraints in robotic learning. Website: zh1hao.wang/Robo_MUTUAL

1 INTRODUCTION

Developing robots that can understand task specifications from diverse modalities (e.g., image, video, text, speech) is a pivotal research area in robot learning [55, 53, 49]. This not only enhances robot performance but also enriches the human-robot interaction experience [23, 13, 42, 62]. To correctly interpret multimodal task specifications, one essential capability required is *Cross-modality Alignment* [55, 23, 62, 47], where an integral high-level task goal exists across various modalities of instructions (or known as prompts [55, 23]) to prevent confusion. Existing methods typically acquire this ability through extensive end-to-end training, raising high demand for meticulously annotated multimodal prompts [49, 55, 53]. Collecting such prompts via crowd-sourcing is notably expensive and laborious [55, 61], not to mention impractical, when only unimodal prompts are accessible (*e.g.*, text instructions are absent and only visual goals are provided). There have been attempts to synthetically generate missing prompts, but how to ensure data quality remains an open question [61]. Therefore, we wonder " *Can we bypass the stringent demands for paired multimodal prompts via unimodal task learning?*", overcoming significant data constraints and opening new avenues for efficient robots learning.

[*]Equal contribution
[†]Corresponding author

Preprint. Under review.

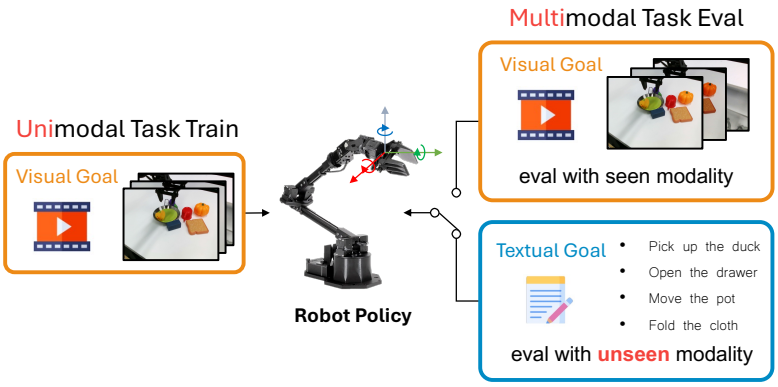


Figure 1: Training robot policies on unimodal task prompts but evaluate using prompts across multi-modalities.

We posit that a positive answer is achievable if the *Cross-modality Alignment* capability can be pretrained using multimodal encoders [52, 33, 65, 34]. Consider a multimodal encoder capable of producing representations for different modalities of prompts that share an identical high-level task goal. Can we find an effective way to encode the textual and visual embeddings of prompts that describe the same task (e.g. "open the door") in a unified representation space [20]? If so, prompts across different modalities will become interchangeable in this shared latent space, allowing unimodal data to implicitly serve as a proxy for multimodal data, which have been demonstrated possible in other multimodal domains like text-to-image generation and different modalities are converging in representation spaces [20, 65, 32, 34].

However, two main challenges must be addressed to apply this methodology to robots: *1)* Existing multimodal encoders are not directly applicable in our setting. They either are not tailored for robot learning [52], or trained solely on narrow scopes of human activity data [14, 10, 15], different from robotics domain [33, 38, 45, 25]. *2)* Even with well-adapted multimodal encoders, a *modality gap* between the representations of different modalities still persists [65, 32, 34], preventing the convergence of different modalities into a consistent latent space.

We propose *Robo-MUTUAL* (Robotic MUltimodal Task specifications via UnimodAl Learning). This new framework enhances the *Cross-modality Alignment* capability of existing multimodal encoders by consuming a broader spectrum of robot-relevant data. Specifically, we retrain DecisionNCE [33], a state-of-the-art robotic multimodal encoder on an all-encompassing dataset, which not only consists large-scale robot datasets including Open-X [50] and DROID [27], but also incorporates a large human-activity dataset EPICK-KITCHEN [10]. Combined, these datasets form the most comprehensive collection to date for robotic multimodal encoder pretraining. Building on the pretrained encoders, we explore two training-free methods to bridge the *modality gap* within the representation space, where we further introduce an effective cosine-similarity noise to facilitate efficient data augmentation in representation space to enable generalization to new task prompts. Tested across over 130 tasks and 4000 evaluations on both simulated LIBERO [35] environments and the real robots platforms, extensive experiments showcase a promising avenue towards enabling robots to understand multimodal instructions via unimodal training.

2 RELATED WORK

2.1 Task Specification in Robot Learning

A series of task specifications have been extensively explored in robot learning. Some works utilize flexible specifications across various modalities to specify the tasks, such as language instructions [4, 5, 13, 49, 28, 33, 38, 25, 60, 1, 56] or visual goals [39, 46, 8, 9, 63, 3, 21], but are specialized for unimodal task specifications. In recent years, some works explore multimodal task specifications [37, 42, 40, 22, 55, 23, 62] and demonstrate positive benefits of leveraging multimodal prompts to enhance the robot policies. In these methods, *Cross-modality Alignment* is one critical capability for multimodal task specifications, where multimodal prompts that share the same high-level task goal should be encoded as similar representations to encourage a well-organized representation

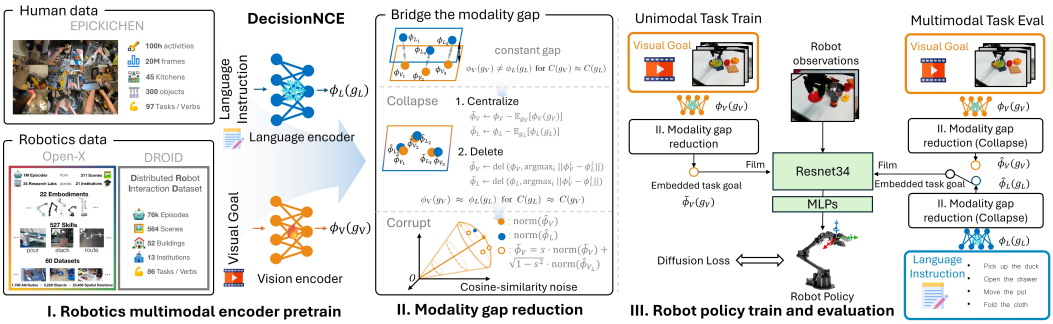


Figure 2: Robo-MUTUAL training pipeline. *I*. Pretrain robotic multimodal encoder consuming broader out-of-domain human and robotics data. *II*. Utilize the pretrained powerful *Cross-modality Alignment* capability and further bridge the modality gap in an efficient and training-free manner. *III*. Achieve multimodal task specifications via unimodal task learning leveraging the well-aligned multimodal representations.

space [55, 40, 42, 62]. However, these methods typically train such ability from scratch on limited robot data with carefully annotated or synthetic multimodal prompts. In this paper, instead, we aim to bypass the restrictive demands on paired multimodal prompts but directly utilize unimodal prompts to achieve multimodal task specifications utilizing the powerful pretrained mutlimodal encoders.

2.2 Multimodal Representation in Robot Learning

Numerous multimodal encoders, such as CLIP [52], BLIP [29] and BLIP2 [31], are designed to align various modalities within a unified representation space, demonstrating notable success in various areas such as image/video caption [41, 18, 58], text-to-image generation [69], and also robotics [26]. However, these encoders are not optimally suited for robot learning as they often fail to capture the temporal visual dynamics critical for robotics [39]. To address this shortfall, specialized robotic multimodal encoders, such as R3M [45], LIV [38], Voltron [25], Lorel [44] and the recent SOTA DecisionNCE [33], have been developed to extract these robotic-critical features. However, the *Cross-modality Alignment* ability of these models is constrained by the narrow scope of their training data, typically limited to specific human datasets [10, 15, 14], without covering diverse out-of-domain robotic data [27, 59, 10, 50].

2.3 Modality Gap in Multimodal Representations

Human can easily summarize similar mental thoughts from multimodal prompts [67, 11, 6], which is also noticed in neural networks [20] where multimodal representations with same semantics are converging. This shows the potential for multimodal encoders to construct a well-aligned representation space, where modalities are interchangeable and unimodal data can approximate multimodal information. However, a persistent modality gap remains even with well-trained multimodal encoders, preventing the convergence of different modality features [34]. Several methods such C3 [65], CapDec [48], and LAFITE [69], etc [54] try to bridge this gap, but the efficacy of these methods and how to enhance the generalization for robotics remain unexplored.

3 Method

3.1 Problem Formulation

We aim to learn a goal-conditioned policy $\pi_\theta(a|s, g)$ that observes state $s \in \mathcal{S}$ and outputs actions $a \in \mathcal{A}$ conditioned on a task prompt g in various modalities, using a small in-domain robot dataset $\mathcal{D}_I = \{(s_i, a_i, g_i)\}_{i=1}^N$. In this paper, we study the most popular prompt g modalities including free-form language instruction $g_L \in \mathcal{G}_L$ and fine-grained visual goal $g_V \in \mathcal{G}_V$ [37, 42, 40, 22], *i.e.*, $g \in \{g_V, g_L\}$ and will explore more modalities like audio in future work. We assume the prompts g in dataset $\mathcal{D}_I$ are unimodal. For instance, in the case of g_L is missing, we can only obtain a unimodal visual-goal conditioned policy $\pi_\theta(a|s, g_V)$ in common sense. In this paper, we also hope to let π_θ

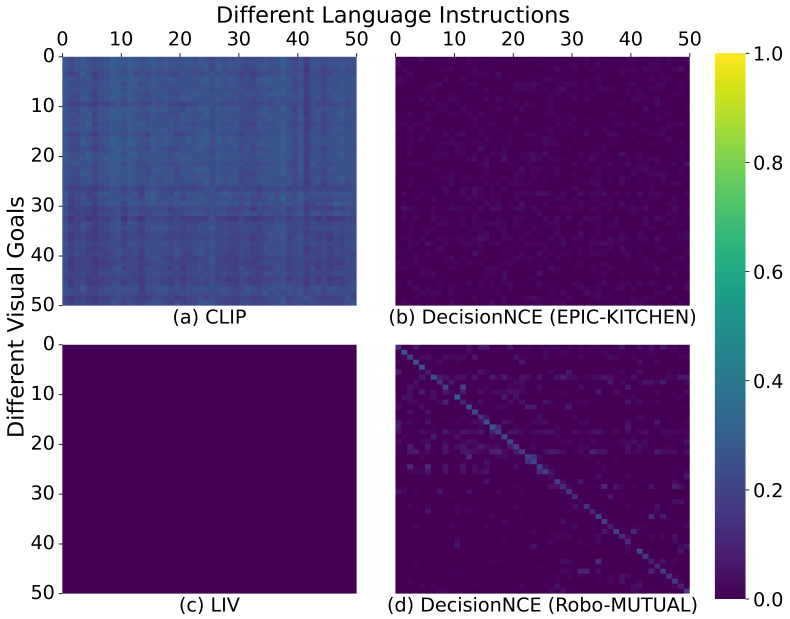


Figure 3: Heatmaps of cosine similarity between representations of language and visual goals. The diagonals are matched pairs. DecisionNCE (Robo-MUTUAL) enjoys strong *Cross-modality Alignment* capability after absorbing broader out-of-domain data.

understand language prompts and execute similar behaviours given g_L that depicts the same task as g_V, *i.e.*, $d^{\pi_\theta}_{g_L}(s) \approx d^{\pi_\theta}_{g_V}(s), \forall s \in \mathcal{S}$ for $C(g_L) \approx C(g_V)$, where $C(g)$ is the high-level task described by prompts g across different modalities, $d^{\pi_\theta}_{g_L}$ and $d^{\pi_\theta}_{g_V}$ denote the state occupancy, or widely known as visitation distribution [43, 30], following the policy $\pi_\theta(a|s, g_L)$ and $\pi_\theta(a|s, g_V)$, respectively.

3.2 Cross-modality Alignment in Representation Space

In standard training process, mostly, $d^{\pi_\theta}_{g_L}(s) \neq d^{\pi_\theta}_{g_V}(s)$, since language instructions g_L and visual goals g_V are two distinct modalities, *i.e.*, $g_L \neq g_V$, where the central challenge is *Cross-modality Alignment*, that similar high-level tasks $C(g_L) \approx C(g_V)$ should be extracted by the policy π_θ to avoid this conflicts [55, 23, 62, 47, 40, 22].

We aim to achieve strong *Cross-modality Alignment* ability by utilizing powerful multimodal encoders (ϕ_V, ϕ_L) pretrained on a diverse out-of-domain dataset $\mathcal{D}_O$ instead of training from scratch based on the limited in-domain robot data $\mathcal{D}_I$ like previous works [55, 22, 40, 47]. If such encoders are accessible, multimodal prompts that encode similar tasks can be projected as interchangeable representations, *i.e.*, $\phi_V(g_V) \approx \phi_L(g_L)$ for $C(g_L) \approx C(g_V)$. Then, we can train multimodal policies π_θ capable of understanding prompts across various modalities within this unified representation space trained solely on unimodal prompts.

$$\begin{aligned} C(g_V) \approx C(g_L) &\Leftrightarrow \phi_V(g_V) \approx \phi_L(g_L) \\ &\Leftrightarrow d^{\pi_{\theta,\phi_L}}_{g_L}(s) \approx d^{\pi_{\theta,\phi_V}}_{g_V}(s), \forall s \in \mathcal{S}. \end{aligned} \tag{1}$$

where π_{θ,ϕ_L}, π_{θ,ϕ_V} denotes $\pi_\theta(a|s, \phi_L(g_L))$ and $\pi_\theta(a|s, \phi_V(g_V))$, respectively. Targeting this goal, we propose *Robo-MUTUAL* (Robotic MUltimodal Task specifications via UnimodAl Learning), containing three parts (Fig 2): *Robotic Multimodal Encoder Pretrain* (Section 3.3), *Modality Gap Reduction* (Section 3.4), and *Robot Policy Train and Evaluation* 3.5).

3.3 Robotic Multimodal Encoder Pretraining

Numerous multimodal encoders are available [52, 38, 33], but are not specifically designed for robotics or trained solely on limited human video that fails to cover the diverse robotics domain. We evaluate the *Cross-modality Alignment* capability of several popular multimodal encoders including

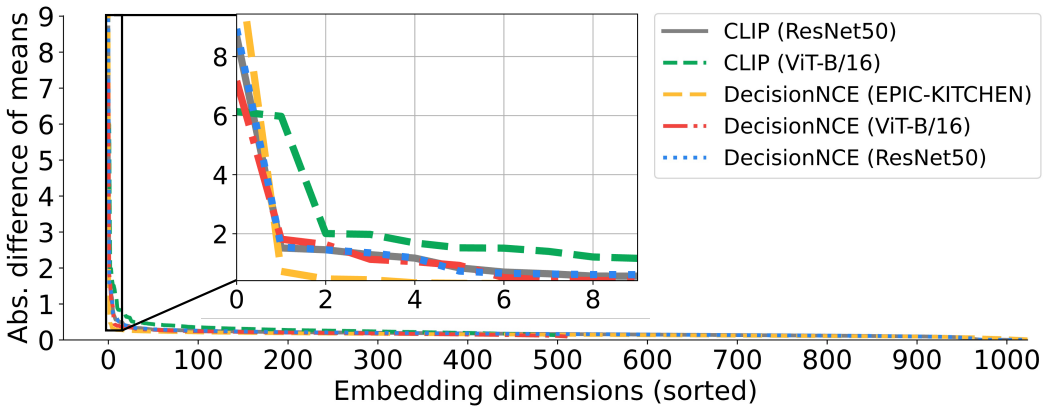


Figure 4: Abs. difference in the means of each embedding dimension cross different modalities. The modality gap manifests in a few dimensions with large discrepancies across modalities, while others remain consistent.

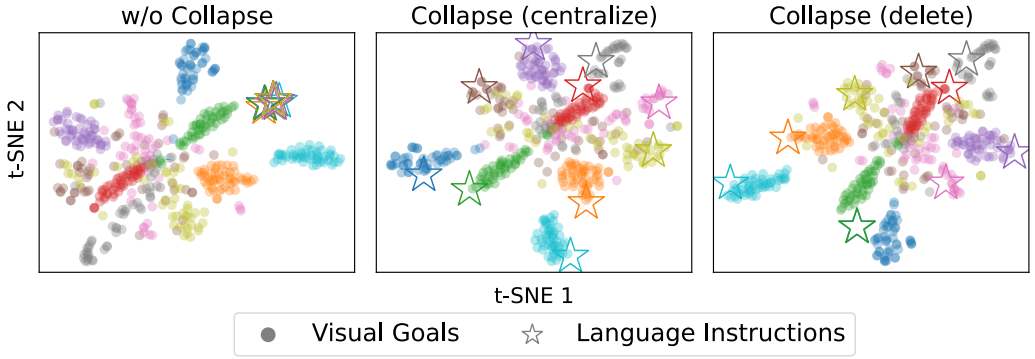


Figure 5: t-SNE [57] projection of DecisionNCE (Robo-MUTUAL) representations of matched visual and language goals. Although different modalities are separate initially, indicating a huge modality gap, this gap can be reduced through the simple *centralize* or *delete* Collapse methods.

CLIP [52], LIV [38] and DecisionNCE [33]. Fig 3 shows that these encoders fail to align matched visual goals and language instructions well in the robotics domain.

Hence, we aggregate diverse large-scale robot-relevant data for robotic multimodal encoder training, aiming at improve the *Cross-modality Alignment* capability as much as possible. This dataset includes Open-X dataset [50], DROID [27], and EPICK-KITCHEN [10], forming a comprehensive dataset $\mathcal{D}_O$ that spans diverse skills/tasks and scenarios. Based on this dataset, numerous robotic multimodal encoders could be viable, we opted to retrain the recent SOTA DecisionNCE [33] for its superior temporal consistency and less sensitivity for hyper-parameter tuning. In difference, we freeze the language encoder ϕ_L from the pretrained CLIP [52] model, focusing solely on contrasting visual goals, as we observe CLIP [52] exhibits robust textual generalization, which is likely attributed to the extensive language data used in CLIP pretraining compared to that in robotics domains. We denote our encoder as DecisionNCE (Robo-MUTUAL) and the training objective is as follows:

$$\min_{\phi_V} \frac{1}{B} \sum\nolimits_{i=1}^{B} -\log \frac{\exp \mathcal{S}(\phi_V(o_{n_i+m_i}) - \phi_V(o_{n_i}), \phi_L(l_i))}{\sum_{j=1}^{B} \exp \mathcal{S}(\phi_V(o_{n_j+m_j}) - \phi_V(o_{n_j}), \phi_L(l_i))}, \tag{2}$$

where n is a random selected start frame in a video clip, m is a random segmentation length and B is batch size. We set $B = 1024$ and train it on 8×A100 GPU for 8 days. See from Fig 3 that DecisionNCE (Robo-MUTUAL) aligns matched visual and textual goals well, where $\mathcal{S}(\phi_V(g_V), \phi_L(g_L))$ is high for $C(g_V) \approx C(g_L)$ and $\mathcal{S}$ is cosine similarity. Considering this superior ability, We will release our check point for DecisionNCE (Robo-MUTUAL) to support researchers develop other future applications conveniently.

3.4 Modality Gap Reduction

Nevertheless, huge modality gap exists in the pretrained representation space, where different modalities with similar semantic meanings fail to absolutely converge but instead are clustered per modality, resulting in $\phi_V(g_V) \neq \phi_L(g_L)$ for $C(g_V) \approx C(g_L)$ [65, 24], as shown in Fig 5. Inspired by recent progresses in minimizing the modality gaps [65, 32, 34, 64], we apply two simple yet efficient *Collapse* and *Corrupt* manipulation to fill the gap.

Collapse. Specifically, the modality gap is primarily characterized as a constant gap between the span of the visual and textual representations clusters $e_V := \{\phi_V(g_V)\}, g_V \sim \mathcal{G}_V$ and $e_L := \{\phi_L(g_L)\}, g_L \sim \mathcal{G}_L$, as proved in [65, 34]. This gap manifests in a few dimensions of the embeddings, which display significant discrepancies across modalities, while the remaining dimensions remain consistent, as shown in [54] and is also observed in Fig 4. Thus, intuitively this gap can be straightforwardly removed by deleting the most different dimensions to address the modality gap [54].

$$\begin{aligned} \text{detele}: \hat{\phi_V} &\leftarrow \text{del}(\phi_V, \arg\max_i \|\phi_V^i - \phi_L^i\|), \\ \hat{\phi_L} &\leftarrow \text{del}(\phi_L, \arg\max_i \|\phi_V^i - \phi_L^i\|), \end{aligned} \tag{3}$$

where, ϕ^i denotes the i-th dimension of ϕ, and $\text{del}(\phi, i)$ denotes deleting the i-th dimension of ϕ. Or we can reduce each modality by its mean values to consider the modality difference across all dimensions, as proved in [64, 34, 65]:

$$\begin{aligned} \text{centralize}: \hat{\phi_V} &\leftarrow \phi_V - \mathbb{E}_{g_V}\left[\phi_V(g_V)\right], \\ \hat{\phi_L} &\leftarrow \phi_L - \mathbb{E}_{g_L}\left[\phi_L(g_L)\right], \end{aligned} \tag{4}$$

where we approximate the expectation term via mini-batch samples from the dataset $\mathcal{D}_O$. After these simple manipulation, the collapsed multimodal representations can mostly bridge the modality gap and collapse together, *i.e.*, $\hat{\phi_V}(g_V) \approx \hat{\phi_L}(g_L)$ for $C(g_V) \approx C(g_L)$, as shown in Fig 5. In our paper, we report the main results using *Centralize* by default and provide ablation studies in experiments.

Corrupt. Note that one language can describe diverse visual goals and vice versa. Hence, we can conveniently augment the collapsed representations to improve the generalization by adding some random noise, which enables π_θ potentially understand unseen prompts. Some works argue that the simple Gaussian noise is effective [65], but we found the cosine similarity noise offers superior augmentation in the high-dimensional embedding spaces [36]. Here we slightly abuse the notation by simplifying ϕ_V or ϕ_L as ϕ:

$$\tilde{\phi} \leftarrow s \cdot \text{norm}(\hat{\phi}) + \sqrt{1-s^2} \cdot \text{norm}(\phi_\perp), \hat{\phi}_\perp = v - \frac{v \cdot \hat{\phi}}{\hat{\phi} \cdot \hat{\phi}} \cdot \hat{\phi} \tag{5}$$

where v is a random embedding, $\hat{\phi}_\perp$ is the orthogonal vector in v *w.r.t* $\hat{\phi}$, $\text{norm}(\text{x}) := \frac{\text{x}}{\|\text{x}\|}$, s is a cosine similarity randomly selected from $[\alpha, 1]$ and we set α to 0.2 as default. The direction of augmented representations remain close to the original to preserve the semantics after corruption, *i.e.*, in (5), $\mathcal{S}(\tilde{\phi}, \hat{\phi}) = \frac{\tilde{\phi} \cdot \hat{\phi}}{\|\tilde{\phi}\| \|\hat{\phi}\|} = s \geq \alpha$, which however can be potentially destroyed by a too large Gaussian noise.

3.5 Robot Policy Training and Evaluation

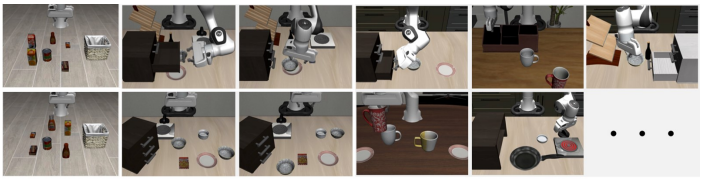
Now, we can achieve *Cross-modality Alignment* in representation space. For instance, in scenarios where only visual goals g_V are available and language instructions g_L are absent, we can utilize the pretrained DecisionNCE (Robo-MUTUAL) visual encoder ϕ_V to generate the visual goal representations $\phi_V(g_V)$, then derive the final corrupted representations $\tilde{\phi_V}(g_V)$ for training the robot policy π_θ. During deployment, the policy π_θ can be prompted with either visual goals g_V or language goals g_L, by using the collapsed and aligned representations $\hat{\phi_V}(g_V)$ or $\hat{\phi_L}(g_L)$.

In training details, we utilize ResNet34 [17] to extract visual feature from both a base and wrist view, where task embedding ϕ is injected via Film conditioning layers [51]. Then, the visual feature is passed through a residual MLPs to predict actions similar to IDQL [16]. The policy is optimized with diffusion loss [19] for its superior effectiveness to model complex distributions [7, 68, 59]. We also use action chunking [66] to improve the policy smoothness.

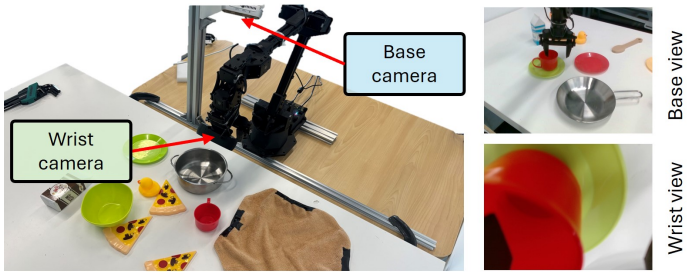
4 Experiments

The experiments try to answer the following questions:

- Can Robo-MUTUAL achieve multimodal task specifications with unimodal data training?
- Does Robo-MUTUAL outperform baseline models that use synthetic prompts generated from unimodal data for multimodal task specifications?
- How do specific design choices for Robo-MUTUAL, such as the use of enhanced robotic multimodal encoders, various collapse, corruption methods, and different scales of corruption, contribute to its effectiveness?



(a) LIBERO task visualizations



(b) Real robot hardware setup

Figure 6: Simulation and real world robotics evaluation setups.

4.1 Experimental Setup

Simulation Environments. We employ the LIBERO benchmark [35], which consists of 130 robotic simulation tasks across four distinct suites: LIBERO-Goal/Object/Spatial/100. LIBERO-100 contains 100 tasks, while the other three suites each have 10 tasks. Each task is accompanied by a language instruction g_L and 50 expert trajectories. Collectively, these suites contain 6500 expert demonstrations and 130 different tasks, which we refer to as LIBERO-All in this paper.

Real Robot Environments. We evaluate 6 tasks on a WidowX robot, spanning skills include `pick & place`, `fold`, `flip` and `move`. We collect around 100 demonstrations per task using the Bridgedata system [59], and annotate each task with a language instruction g_L. Successful catching correct object is recorded as half-completed of the whole task, and a continued correct placement will scored a full success.

Evaluation Scenarios. *1)* We train Robo-MUTUAL in a common scenario where only visual goals g_V are available. Visual goal is provided as the transition between the representations of the initial and final frames of a video clip [33]. In this case, the policy only observes visual goals g_V during training but is required to understand both visual goals g_V and language instructions g_L during evaluation. *2)* We also explore the reverse scenario, *i.e.*, training exclusively on textual goals g_L but evaluating with visual goals g_V to assess bidirectional transferability across modalities. Due to the high cost associated with real-world evaluations, we compare Robo-MUTUAL against baseline methods in simulations. However, we believe the 130 tasks in LIBERO benchmark offer comprehensive coverage across a diverse range of tasks and skills, ensuring fair and sufficient comparisons.

Baselines. 1) GPT4-Synthetic is the main baseline in the common scenario that textual goals are missing. This approach represents a series of methods that firstly generate missing language instructions from unimodal visual goals using pretrained large multimodal models [61, 2], and then

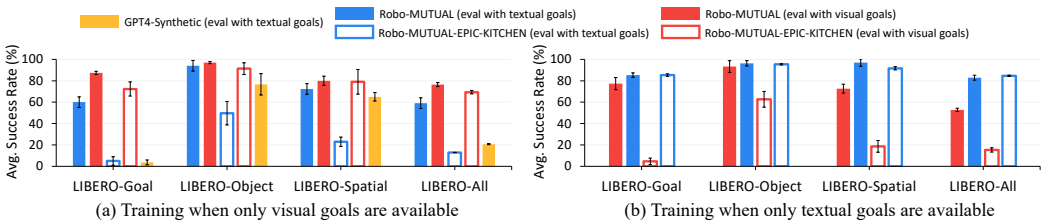


Figure 7: **Simulation evaluation.** (eval with textual/visual goals) denote the robot policy is evaluated with textual goal and visual goals, respectively. For each bar, the robot policy is trained on 3 different random seeds and is evaluated for 10 episodes for each task.

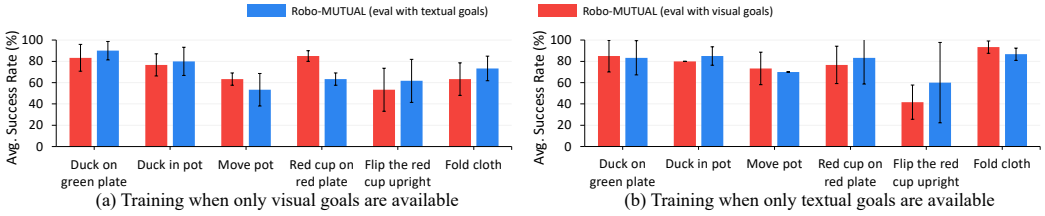


Figure 8: **Real robot evaluation.** (eval with textual/visual goals) denote the robot policy is evaluated with textual goal and visual goals, respectively. For each bar, the robot policy is trained on 3 different random seeds and is evaluated for 10 episodes for each task.

train policies conditioned on these synthetic goals. During evaluation, we prompt the policies using the ground truth language instructions to evaluate the success rates. To ensure a fair comparison, we employ the advanced GPT-4V to generate language instructions based on the initial and final frames of a trajectory. Furthermore, we incorporate several ground truth examples to enhance the quality of synthetic instructions leveraging the substantial in-context learning capability of large models [12].
2) Robo-MUTUAL-EPICK adheres the same implementation protocols as our Robo-MUTUAL, but directly uses the DecisionNCE [33] pretrained solely on the narrow EPICK-KITCHEN dataset [10] as the robotic multimodal encoder without the specialized improvement in Section 3.3, which may suffer from limited *Cross-modality Alignment* capability.

4.2 Main Results

Transfer from Visual to Textual Goals. In the common scenario where only visual goals are available, both the simulated and real-world evaluations in Fig 7 (a) and Fig 8 (a) clearly demonstrate that Robo-MUTUAL can successfully understand both visual and textual goals trained exclusively on visual goals. Specifically, we only observe a moderate performance drop when evaluating Robo-MUTUAL from visual to textual goals. However, this property is not enjoyed by Robo-MUTUAL-EPICK, which fails to transfer from visual to textual goals and undergoes severe performance drop, as shown in Fig 7 (a), primarily due to the limited *Cross-modality Alignment* capability (see Fig 3 for details). Moreover, GPT4-Synthetic can partially work well using the synthetic language instructions thanks to the superior capability of large models, however, still underperforms Robo-MUTUAL. We think this is because GPT-4V only consider limited robotic data during pretraining and it is hard to guarantee the quality of synthetic data [61].

Transfer from Textual to Visual Goals. We also explore the reverse scenario in which Robo-MUTUAL is trained exclusively on textual goals and then evaluated with visual goals. In Fig 7 (b) and Fig 8 (b), Robo-MUTUAL demonstrates effective transferability in this reversed setup. Importantly, the visual goals provided for task specifications do not align precisely with the robot's current observations, *i.e.*, discrepancies exist in the location of target object, the target position, and even the number, location, types of distractors. Under this hard condition, Robo-MUTUAL must accurately extract correct semantic task information from visual goals that contain many distracting elements in a zero-shot manner trained solely with textual goals. Overall, Robot-MUTUAL demonstrates strong bidirectional transferability to enable multimodal task instructions via unimodal learning*.

*See detailed videos in `zh1hao.wang/Robo_MUTUAL`.

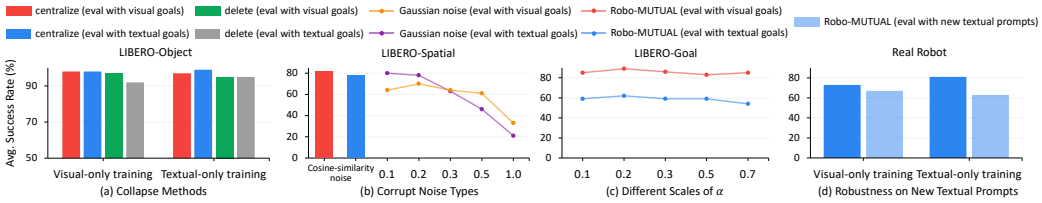


Figure 9: **Ablation studies.** (a) Comparison between (*delete* and *centralize*) collapse methods; (b) Comparison between *Cosine-similarity noise* and Gaussian noise; (c) Corrupt strength of *Cosine-similarity noise*; (d) The generalization ability on unseen new textual instructions

4.3 Ablation Studies

Collapse Methods. In the main results, we only report *centralize* due to space limits. Here, we compare the *centralize* and *delete* collapse methods. Fig 9(a) shows that they achieve comparable performance and enjoy moderate performance gap. Also, Fig 5 shows that both of them effectively cluster matched textual and visual pairs. One future work is to investigate the conditions under which *centralize* and *delete* outperform each other in the robotics domain.

Corrupt Noise Types. We compare our *Cosine-similarity noise* with the widely adopted simple Gaussian noise with different standard deviation (Std) [65]. Fig 9 (b) demonstrates that the simple Gaussian noise is quite unstable *w.r.t* different Std and is inferior to *Cosine-similarity noise*. Intuitively, this is because *Cosine-similarity noise* naturally enables the augmented data to preserve the semantics of its origins, as the augmented data remains a high cosine similarity with the original data, *i.e.*, $\mathcal{S}(\tilde{\phi}, \hat{\phi}) \in [\alpha, 1]$. However, this can be easily violated by the simple Gaussian noise, as a too large Gaussian noise can mostly reverse the direction of the original representation after augmentation and thus lost its original semantics. On the other hand, however, a too small noise lacks enough augmentations, making it quite sensitive to tune the Std of the Gaussian noise.

Corrupt Strength. To further demonstrate the effectiveness of *Cosine-similarity noise*, we ablate on different corrupt strength by adjusting the cosine-similarity threshold α. Fig 9 (c) shows that *Cosine-similarity noise* is robust to various corrupt strength, maintaining consistent performance across various α. Meanwhile, $\alpha = 0.2$ works best in our evaluation and is set as our default choice. This likely correlates with the pre-corruption cosine-similarity of matched pairs, which averages around 0.2, as shown in Fig 3(d).

Robustness on New Textual Prompts. We also investigate whether Robot-MUTUAL can generalize beyond the ground-truth goals in the downstream dataset $\mathcal{D}_I$. For example, we replace the original "pick up the red cup and place it on the plate" as "move the red cup to the plate", sharing the same abstracted task goal but are expressed differently. Fig 9 (d) shows that Robo-MUTUAL enjoys robustness to such new textual prompts thanks to the strong *Cross-modality Alignment* capability that maps different instructions with similar semantics as similar representations. This means that we can conveniently augment instructions in the latent space in one implicit way by simply adding noise, like the *Cosine-similarity noise*, rather than relying on heavy large models to explicitly synthesize or refine instructions like GPT4-Synthetic and other relevant works [61, 2].

5 Conclusions

We introduce Robo-MUTUAL, a framework that enbales robots to comprehend multimodal task specifications using only unimodal prompts. This is achieved by treating multimodal instructions as interchangeable embeddings within a well-aligned multimodal representation space, leveraging the strong *Cross-modality Alignment* capability from pretrained encoders on a comprehensive robotic dataset and two simple yet effective modality gap reduction methods. Extensive evaluations on both real and simulated robots validate the effectiveness of our approach. One limitation is we only consider the language and image modalities, and we will explore more modalities like audio in future work. Further enhancements will also focus on refining multimodal encoders and improving modality gap reduction techniques.

References

[1] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024.

[2] Nils Blank, Moritz Reuss, Fabian Wenzel, Oier Mees, and Rudolf Lioutikov. Scaling robot policy learning via zero-shot labeling with foundation models. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024.

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

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[6] David Burr and Monica Gori. Multisensory integration develops late in humans. 2012.

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[8] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind 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] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. In *The Eleventh International Conference on Learning Representations*, 2023.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[11] Solange Denervaud, Edouard Gentaz, Pawel J Matusz, and Micah M Murray. Multisensory gains in simple detection predict global cognition in schoolchildren. *Scientific reports*, 10(1):1394, 2020.

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[13] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu 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.

[16] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[20] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[21] Stephen James, Michael Bloesch, and Andrew J Davison. Task-embedded control networks for few-shot imitation learning. In *Conference on robot learning*, pages 783–795. PMLR, 2018.

[22] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[23] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In *Fortieth International Conference on Machine Learning*, 2023.

[24] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

[25] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

[26] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022.

[27] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[30] Jianxiong Li, Xiao Hu, Haoran Xu, Jingjing Liu, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. Mind the gap: Offline policy optimization for imperfect rewards. In *The Eleventh International Conference on Learning Representations*, 2023.

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[32] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023.

[33] Jianxiong Li, Jinliang Zheng, Yinan Zheng, Liyuan Mao, Xiao Hu, Sijie Cheng, Haoyi Niu, Jihao Liu, Yu Liu, Jingjing Liu, et al. Decisionnce: Embodied multimodal representations via implicit preference learning. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[35] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[36] Yang Liu, Xiaomin Yu, Gongyu Zhang, Christos Bergeles, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Arcsin: Adaptive ranged cosine similarity injected noise for language-driven visual tasks. *arXiv preprint arXiv:2402.17298*, 2024.

[37] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[38] Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pages 23301–23320. PMLR, 2023.

[39] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

[40] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022.

[41] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[42] Vivek Myers, Andre Wang He, Kuan Fang, Homer Rich Walke, Philippe Hansen-Estruch, Ching-An Cheng, Mihai Jalobeanu, Andrey Kolobov, Anca Dragan, and Sergey Levine. Goal representations for instruction following: A semi-supervised language interface to control. In *Conference on Robot Learning*, pages 3894–3908. PMLR, 2023.

[43] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

[44] Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea 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.

[45] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[46] Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. *Advances in neural information processing systems*, 32, 2019.

[47] Andre T Nguyen, Luke E Richards, Gaoussou Youssouf Kebe, Edward Raff, Kasra Darvish, Frank Ferraro, and Cynthia Matuszek. Practical cross-modal manifold alignment for robotic grounded language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1613–1622, 2021.

[48] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022.

[49] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[50] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[51] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[53] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[54] Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

[55] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multimodal task specifications. In *7th Annual Conference on Robot Learning*, 2023.

[56] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv:2403.12910*, 2024.

[57] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[58] Lucas Ventura, Cordelia Schmid, and Gül Varol. Learning text-to-video retrieval from image captioning. *arXiv preprint arXiv:2404.17498*, 2024.

[59] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[60] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Robotic skill acquisition via instruction augmentation with vision-language models. *arXiv preprint arXiv:2211.11736*, 2022.

[61] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Robotic skill acquistion via instruction augmentation with vision-language models. In *Proceedings of Robotics: Science and Systems*, 2023.

[62] Albert Yu and Ray Mooney. Using both demonstrations and language instructions to efficiently learn robotic tasks. In *The Eleventh International Conference on Learning Representations*, 2023.

[63] Tianhe Yu, Chelsea Finn, Sudeep Dasari, Annie Xie, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *Robotics: Science and Systems XIV*, 2018.

[64] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations (ICLR)*, 2023.

[65] Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. *arXiv preprint arXiv:2401.08567*, 2024.

[66] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[67] Bingxin Zhao, Yujue Li, Zirui Fan, Zhenyi Wu, Juan Shu, Xiaochen Yang, Yilin Yang, Xifeng Wang, Bingxuan Li, Xiyao Wang, et al. Eye-brain connections revealed by multimodal retinal and brain imaging genetics. *Nature Communications*, 15(1):6064, 2024.

[68] Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. *arXiv preprint arXiv:2401.10700*, 2024.

[69] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17886–17896, 2022.