Any2Policy: Learning Visuomotor Policy with Any-Modality

Yichen Zhu, Zhicai Ou, Feifei Feng, Jian Tang[*]
Midea Group

Abstract

Humans can communicate and observe media with different modalities, such as texts, sounds, and images. For robots to be more generalizable embodied agents, they should be capable of following instructions and perceiving the world with adaptation to diverse modalities. Current robotic learning methodologies often focus on single-modal task specification and observation, thereby limiting their ability to process rich multi-modal information. Addressing this limitation, we present an end-to-end general-purpose multi-modal system named Any-to-Policy Embodied Agents. This system empowers robots to handle tasks using various modalities, whether in combinations like text-image, audio-image, text-point cloud, or in isolation. Our innovative approach involves training a versatile modality network that adapts to various inputs and connects with policy networks for effective control. Because of the lack of existing multi-modal robotics datasets for evaluation, we assembled a comprehensive real-world dataset encompassing 30 robotic tasks. Each task in this dataset is richly annotated across multiple modalities, providing a robust foundation for assessment. We conducted extensive validation of our proposed unified modality embodied agent using several simulation benchmarks, including Franka Kitchen and Maniskill2, as well as in our real-world settings. Our experiments showcase the promising capability of building embodied agents that can adapt to diverse multi-modal in a unified framework.

1 Introduction

What is the ultimate form of robots? It should possess the ability to both listen and read, learn from demonstrations, and perceive the three-dimensional world with an understanding of time. While recent advancements in robot learning have explored various modalities for task specification and environmental perception – such as text, speech, images, videos, and point clouds – most prior research has approached these modalities as distinct challenges.

An expanding field of research within artificial intelligence, spanning various disciplines, indicates that simultaneous learning across different modalities, such as vision-language [1, 2, 3, 4, 5], video-text [6, 7, 8, 9, 10, 11], and vision-touch [12, 13, 14, 15], leads to the development of more comprehensive and effective representations. The enriched representations derived from this approach lead to a more profound comprehension of each modality on a standalone basis. This multi-faceted concept finds backing in a number of cognitive science and psychology studies [16, 17], which propose that human learning is significantly enhanced when it integrates multiple cues, compared to relying on a single modality in isolation. By adopting this multi-modal approach, which leverages the collective strength of various senses or inputs, we emulate the natural human method of information acquisition and processing, indicating a path toward a more integrative methodology in robotics.

[*] Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-4244

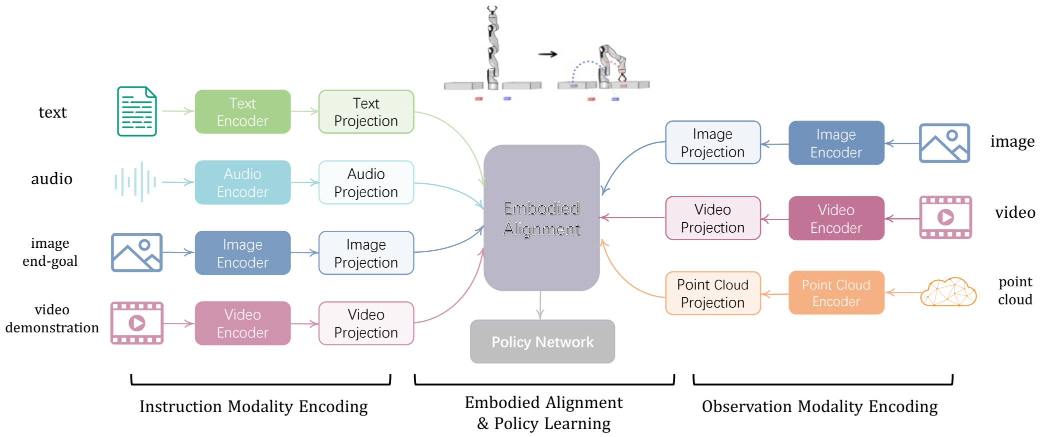

Figure 1: The overall framework of Any2Policy. The Any2Policy framework is structured to handle multi-modal inputs, accommodating them either separately or in tandem at the levels of instruction and observation. We design embodied alignment modules, which are engineered to synchronize features between different modalities, as well as between instructions and observations, ensuring a seamless and effective integration of diverse input types.

Building on the proven success of multi-modal learning in various disciplines, our goal is to develop a multi-modal embodied agent capable of processing a wide array of instruction and observation modalities. Our method, referred to as Any-to-Policy Robot Model (abbreviated as Any2Policy), is illustrated in Figure 1. The Any2Policy architecture is structured in two primary components. Initially, we harness multi-modal encoders to extract relevant features from the input. These features are then transformed into standardized token representations, ensuring uniformity at the neural network level via projection layers. In the second component, we introduce embodied alignment, which processes multi-modal signals, enriched with specific instructions. These signals are routed to encoders following projection, ultimately leading to the generation of actions for manipulation.

In support of this project, we are releasing a substantial real-world dataset consisting of 30 tasks, where each task includes 30 trajectories, all annotated with multi-modal instructions and observations, mirroring the setup used in our experiments. The purpose of this dataset is to foster and encourage future research in the area of multi-modal embodied agents.

We have demonstrated the efficacy of our Any2Policy approach through rigorous testing against this benchmark. The results clearly indicate that Any2Policy is highly adept at interpreting and responding to a variety of modalities. We further conduct experiments on two simulated robotics benchmarks to reinforce the strong generalizability of our approach compared to existing methods. Our project is at any2policy.github.io/.

In summary, our contributions are the follows:

- We introduce any-to-policy models that enable a unified embodied agent to process various combinations of modalities, effectively facilitating instruction and perception of the world.
- We present novel embodied alignment learning techniques designed to seamlessly align instructions and observations, enhancing both the effectiveness and efficiency of policy learning.
- We offer a multi-modal dataset tailored for robotics, encompassing 30 distinct tasks. This dataset covers a wide spectrum of modalities in both instruction and observation.

2 Related Work

Multi-Modal Learning. Recent large vision language models (VLMs) use pretrained image encoders as part of the larger model; some pretrain it with supervised classification, some use the pretrained CLIP encoders [18], and some with custom multi-modal pretraining. Inspired by GPT-4, which has been released and demonstrates many advanced multi-modal abilities, recent models present more lightweight versions of VLM that align the image inputs with a large language model using proper instructional tuning [19, 20, 21, 22]. These vision-language models also showcase many advanced multi-modal capabilities after the alignment. Another line of research [23, 24, 25, 26, 27, 28] seeks to demonstrate the transformer's capability to deal with multiple modalities in a single neural network and then perform image generation, video-question answering, and multi-turn dialogue, graph learning [29, 30, 31] according to the input modalities. Different from these prior works, we focus on robotics. Our design multi-modal input includes both instruction and observation, which potentially makes the robot more generalizable to diverse scenarios.

Embodied agent with Diverse Modalities. The study of embodied AI has been extensively explored through instruction and observation methods. A series of studies concentrate on learning using varied task specifications. This includes language-conditioned policy learning [32, 33, 34, 35, 36, 37, 32, 38, 39, 40], instruction following agents [41, 42] with 3D observation [43, 44, 45, 46, 47], visual goal-reaching [48, 49, 50, 51], video demonstrations [52, 53, 54, 55, 56, 34, 57, 58], or combination with multiple modalities [59, 60, 61]. Notably, most systems are specialized for certain task specifications, with few like VIMA and MUTEX engaging with multiple modalities, albeit focusing on a single observation type.

Representation learning is critical in high-dimensional control settings, particularly when managing visual observation spaces. Recent advancements in this field have explored different modalities in feature learning for robotics, especially for images with the unsupervised pretrained visual representation [62, 63, 64, 65, 66, 67, 68, 69], or using visual encoders from pretrained vision-language models [70, 71, 72, 73, 74, 75, 76, 77] for robotics [78, 79, 80, 33, 81, 82, 59, 83, 84, 61]. These approaches often neglect instruction or rely solely on text-based instruction, lacking exploration of diverse instructional modalities.

Our proposed Any2Policy framework differs from previous work by focusing on both instruction and observation levels and utilizing various modalities. This approach aims to enhance learning and inference for embodied agents, offering a more holistic and integrated perspective compared to earlier, more modality-specific research.

Robot Manipulation and Datasets. A diverse array of robot manipulation tasks necessitates various skills and task specification formats, including instruction following [32], one-shot imitation [85], rearrangement [86], reasoning [87, 88, 89], and planning [90, 91]. Most existing datasets predominantly feature 2D images for observations and, in some cases, include language instructions to prompt specific robotic actions, as seen in datasets like Franka-Kitchen [92] and CALVIN [93]. VIMA-Bench [59] introduces a dataset that combines language instructions with images in multimodal prompts. Meanwhile, Maniskill [94] provides 3D visual information, enabling models to interpret point clouds as observations. Our dataset, RoboAny, stands out as the first to support a comprehensive range of modalities in robotics. It encompasses both instructions and observations across images, videos, audio, language, and point clouds, offering unparalleled diversity in robotic task specification and learning.

3 Methodology

Our objective is to develop a comprehensive policy capable of executing a variety of tasks, utilizing a dataset of demonstrations that are annotated with multi-modal instructions and observational data from various sources such as images, videos, point clouds, text, and speech. We acknowledge that the modality of task specifications and observations may differ depending on the specific scenario, and our approach is tailored to adapt flexibly to these varying modalities.

We start with a formal description of our settings and provide mathematical annotations. For each task $T \in \{T_1, T_2, \cdots, T_n\}$, we assume that the embodied agents learn from demonstrations obtained through teleoperation, $D_i = \{d_i, \cdots, d_m\}$, where m is the number of demonstrations for task T_i. Each demonstration d_i^j presents the form of a sequence of observations and expert actions, $d_i^j = [(s, o, a)_i^j]$, where $a \in A$ denotes the actions, $s \in S$ represent instruction, and $o \in O$ is the observation.

The instruction of task T_i is specified with four modalities: text instructions $s \in \{L_i^1, L_i^2, \cdots, L_i^k\}$, audio instruction $s \in \{A_i^1, A_i^2, \cdots, A_i^k\}$, image goal instruction $s \in \{M_i^1, M_i^2, \cdots, M_i^k\}$ and video demonstration $s \in \{V_i^1, V_i^2, \cdots, V_i^k\}$. Similarly, the observation of task T_i, is detailed by

https://doi.org/10.52202/079017-4244

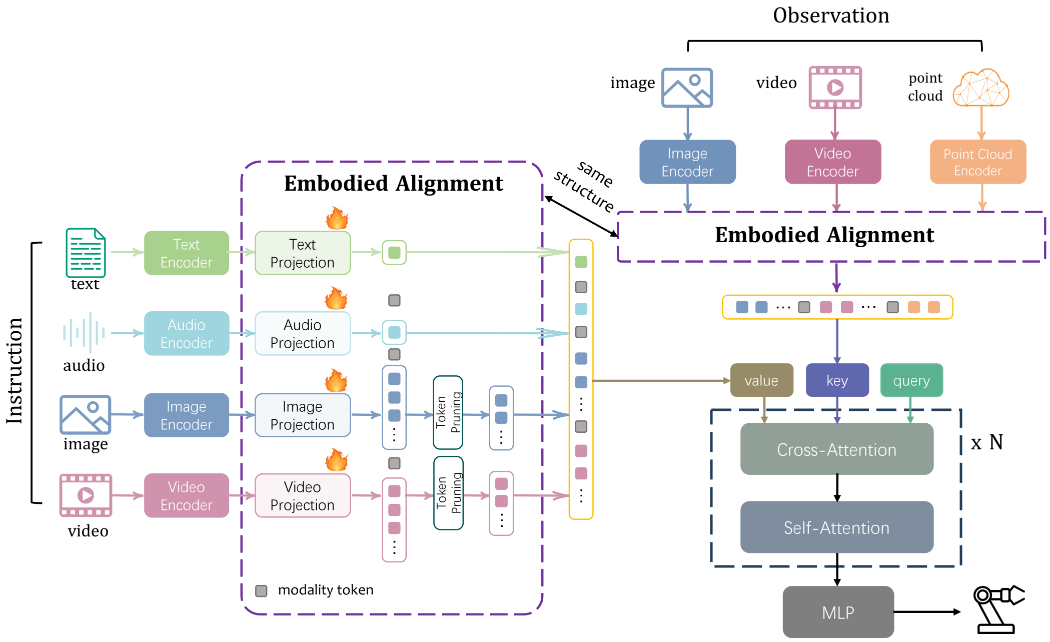

Figure 2: The architecture of embodied alignment and policy network.

m alternative perceiver, each with three modalities: image $o \in \{m_i^1, m_i^2, \cdots, m_i^k\}$, point cloud $o \in \{p_i^1, p_i^2, \cdots, p_i^k\}$, and video $o \in \{v_i^1, v_i^2, \cdots, v_i^k\}$. Capital letters are used to denote instructions, while lowercase letters represent observations.

Our objective is to develop a multi-task policy, denoted as $\pi(a|s, o)$, which generates continuous actions a in response to the current observation o and given instruction s. The policy aims to proficiently execute n tasks included in the training dataset D under new initial conditions, such as varying positions of objects. Moreover, a critical aspect of this policy is its capacity to generalize effectively to environments, entities, and tasks that have not been previously encountered or represented in the training data.

3.1 Overall Architecture

A pivotal characteristic of an Any2Policy agent is its capacity to leverage multi-modal information throughout its training phase, exhibiting the ability to operate competently when restricted to a single modality. It adapts its comprehension and actions to suit the available sensory input. This versatility brings complexities in crafting multi-modal models, requiring careful consideration in their design to ensure seamless adaptability and robust performance. To address these, we introduce two modules: 1) Multimodal encoders to promote cross-modal interactions of all modalities into a shared latent space for instruction and observation, individually, and 2) Embodied alignment to take in the instruction token and merged with observation tokens via cross-attention. We provide an overview in Figure 2.

Multimodal Encoder. There are many choices for the multi-modal encoders, for instance, CLIP [18] for language and vision transformer for images. Specifically, we denote the encoder as a mapping function $P(\cdot)$, and for each modality, we need a particular model to extract useful modality-agnostic representation to prepare for policy learning. To make our framework more consistent, we would like to share the encoder if the instruction and observation have the same modality. Then, we would have five modality encoders for text, audio, image, video, and point cloud, respectively.

Using different backbone models for each modality can be cumbersome and hard to maintain. Thus, we leverage existing well-established models, ImageBind [25], a unified high-performance encoder across five modalities, to encode inputs of various modalities. With the help of ImageBind, we are spared from managing many numbers of heterogeneous modal encoders. Notably, the multi-modal encoders extract semantically meaningful representations from the input modality, which aims to provide essential knowledge to the policy learning networks for accurate manipulation.

https://doi.org/10.52202/079017-4244

Embodied Alignment. We leverage a projection layer to map different modalities' feature representations from $P(x)$ into unified representations. In particular, we want to avoid heavy computational costs during training, so that we freeze all the layers in ImageBind, and only keep the projection layers learnable. This operation is kept the same for both the instruction segment and the observation segment.

As a result, the projection layers ensure that the tensor of different modalities are in the same form. Nevertheless, the process is fraught with challenges, 1) aligning distinct modalities within a segment, particularly when diverse types such as images and point clouds coexist, is difficult. Simultaneously, it's vital to control computational costs, which can escalate with the addition of modalities, and 2) the integration of instructions with observations is complex. This complexity arises partly because instructions and observations don't always correspond one-to-one; often, a single instruction relates to multiple observation frames, and for most time steps, observations exist without accompanying instructions. Additionally, the modality gap between instruction and observation can be an obstacle in creating effective feature representations for manipulation tasks.

To address the first challenge, we employ a transformer architecture. Each modality undergoes tokenization, with the token count varying between modalities. For example, images and videos typically require more tokens to encapsulate visual information. We use 81 tokens for images, $81 \times t$ for videos (where t represents the number of frames), and 256 tokens for point clouds. For text and audio, a single token is generally sufficient due to their shorter length. It's important to note that the computational demands of self-attention layers increase quadratically with the number of tokens, leading to potentially high inference costs for visual modalities. Therefore, it is crucial to use model compression methods [95, 96, 97, 98, 99, 100, 101, 102, 103, 104] to reduce the computational cost.

We follow RT-1 [35] to utilize TokenLearner [105] for pruning redundant visual tokens. We apply TokenLearner individually to each visual modality, such as images, videos, and point clouds. After token pruning, the token count for images and videos is reduced to 8, while for point clouds, it's brought down to 16. These tokens are then amalgamated, using a modality token to distinguish between different modalities. We insert a modality token between every two modalities. We also use absolute position embedding to maintain the token order. Consequently, the maximum token count, when integrating all three visual modalities, is significantly reduced to as few as 34 (8+8+16+2), substantially enhancing computational efficiency. For the transformer model, we use one Transformer block with self-attention and feed-forward networks.

The second major challenge is how to align instruction with observation. We seek the help of cross-attention. Specifically, when we are presented with an observation sequence, P_o, and an instruction sequence, P_s, we project P_o into the key (K) and P_s into the value (V). We create a set number of learnable query (Q) embedding as input for the cross-attention. This learnable query embedding aims to adapt different modalities for efficient fusion between instruction and observation. This setup enables us to align the two tensors through a cross-attention mechanism, mathematically represented as:

$$h_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d_h}} V_i) \tag{1}$$

In this equation, d_h denotes the dimensionality, and the construction of the query, key, and value components aligns with the standard framework utilized in BERT [106]. The elegance of the method lies in its ability to effectively incorporate relevant representations from the instruction into the observation, thus ensuring a more coherent and informative interaction between the two. In our experiments, we use 16 queries, where the dimension of queries corresponds to the cross-attention layers.

Policy Networks. Designing a multi-task policy comes with its share of challenges, and one critical aspect involves choosing an appropriate conditioning mechanism. In our framework, the policy networks are conditioned based on the observation sequence and occasionally an instruction sequence. As aforementioned, this conditioning is achieved through a series of cross-attention layers, which establish connections between the policy network and the input modalities. Each Transformer block comprises self-attention, cross-attention, and a feed-forward network. Additionally, we incorporate residual connections that link the higher layers with the input rollout trajectory sequence. We append one history action token to the model, similar to VIMA approach [59]. It is important to note that the instruction aligns with the observation at every time step. To optimize efficiency, we store the instruction's representation to avoid repetitive computations. The decoder is constructed using L

https://doi.org/10.52202/079017-4244

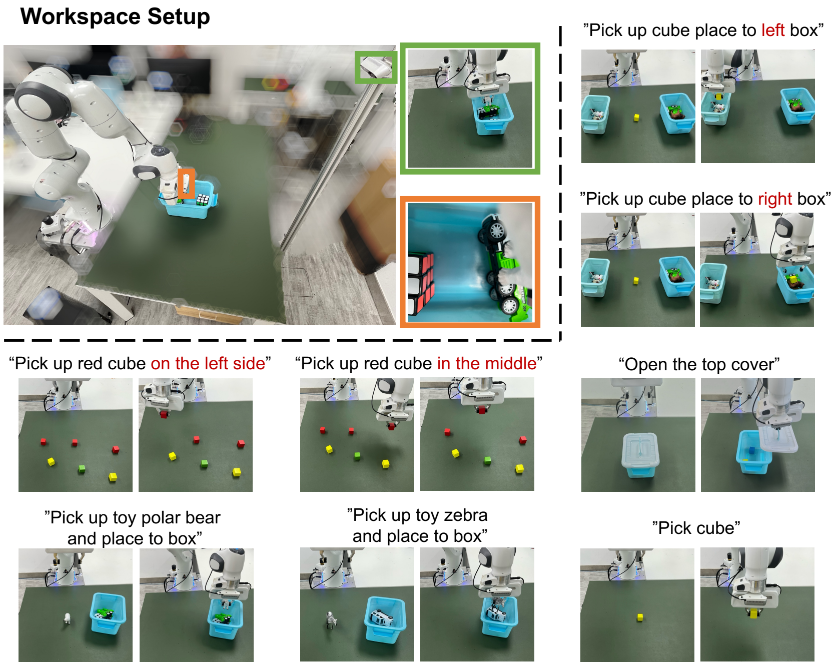
Workspace Setup
"Pick up cube place to left box"
"Pick up cube place to right box"
"Pick up red cube on the left side"
"Pick up red cube in the middle"
"Open the top cover"
"Pick up toy polar bear and place to box"
"Pick up toy zebra and place to box"
"Pick cube"

Figure 3: This is the setup of our Franka real robot. We have compiled examples of several tasks. To facilitate better understanding for our readers, we provide only the language-based versions of these task descriptions.

Transformer blocks. Finally, we append a Multi-layer Perceptron (MLP) layer to generate continuous actions. We use behavior cloning mean squared loss as the optimization objective.

3.2 Dataset Construction

In our pursuit to develop a robot model that allows input with any modality, we have established a new dataset featuring a range of tasks, each accompanied by multiple instructions and observations tailored to real-world scenarios. We developed two workstations, one of which is depicted in Figure 3. This dataset comprises $n = 30$ distinct tasks. These tasks vary from straightforward pick-and-place actions, such as "pick up the yellow cube," and "place the toy bear in the blue box to the right," to more complex contact-rich tasks like "open the drawer" as well as tasks demanding to reason, for instance, "sort the cube with the same color." Although all tasks are executed within the same environment, they involve a variety of objects selected from a set of 70. Each task is exemplified through $m = 30$ human-collected trajectories. Further, every task is annotated with $k = 5$ distinct instructions. To generate these diverse instructions, we utilized GPT-4, creating prompts to formulate alternative descriptions. We have carefully filtered these descriptions to ensure they accurately represent the specific task-relevant objects, avoiding any potential confusion due to synonyms. Additionally, we have incorporated speech signals from various characters of the Amazon Polly service to enrich our dataset with diverse speech descriptions. In terms of observation, we recorded the video during our data collection process. The depth information captured in these recordings was then converted into point cloud data, enabling detailed analysis and application of the collected data. All data are collected via teleoperation. For long-horizon tasks, we operated the robot to move back to the initial state and then start the next move. In the appendix, we provide detailed descriptions of our hardware for two workstations and more qualitative examples of our collected data.

4 Experiments

This section aims to demonstrate the generalizability of our proposed Any2Policy agent. We evaluate our method in two setups. Initially, we conduct experiments in a real-world setting using our own collected dataset, which comprises various modalities. This is aimed at demonstrating the efficacy

https://doi.org/10.52202/079017-4244

Table 1: Comparison with modality-specific models in real-world experiments.

Instruction → Observation	Text → Image	Text → Video	Text → Point Cloud	Audio → Image	Audio → Video	Audio → Point Cloud
Any2Policy	**51**	**57**	**62**	**49**	**55**	**57**
Modality-Specific	39	42	47	26	48	50
Instruction → Observation	Image → Image	Image → Video	Image → Point Cloud	Video → Image	Video → Video	Video → Point Cloud
Any2Policy	**56**	**63**	38	**46**	**57**	**36**
Modality-Specific	45	47	**45**	39	46	28

Table 2: Comparison with Any2Policy framework without embodied alignment in real-world experiments

Instruction → Observation	Text → Image	Text → Video	Text → Point Cloud	Audio → Image	Audio → Video	Audio → Point Cloud
Any2Policy	**51**	**57**	**62**	**49**	**55**	**57**
- Embodied Alignment	26	18	31	27	20	16
Instruction → Observation	Image → Image	Image → Video	Image → Point Cloud	Video → Image	Video → Video	Video → Point Cloud
Any2Policy	**56**	**63**	**38**	**46**	**57**	**36**
- Embodied Alignment	15	29	28	15	9	11

of our approach across a wide range of instruction and observation modalities. Subsequently, we assess the performance of our method on three simulated benchmarks. Each benchmark is limited to one or a small number of modalities. This phase aims to establish that, by leveraging training on cross-modal information, our method can still deliver competitive or superior results compared to standard benchmarks, even in the presence of a single modality.

4.1 Real-World Evaluation

Evaluation Setup. We conduct our evaluations using a newly constructed dataset of multimodal instruction and observation. The dataset is divided into training, validation, and testing subsets, with a split of 7/1/2, respectively. We position the objects randomly, allowing tasks to be performed on unseen objects.

Implementation details. We use an initial learning rate of 3e-5 with the AdamW [107] optimizer, a weight decay of 1e-6, and a linearly decaying learning rate scheduler with a warm-up covering the initial 2% of the total training time [108]. We apply a gradient clipping of 1.0. All experiments are evaluated over 10 trials to obtain the mean success rate. The action space is the absolute joint position.

Experiment Results. We aim to answer the following questions with our experiments.

1. Does Any2Policy outperform single-modality trained methods? We compare Any2Policy with modality-specific approaches in Table 1. In particular, all implementations and network architecture remain constant across both settings, except that Any2Policy is trained on a variety of modalities while the modality-specific method is limited to particular ones. We observe that Any2Policy consistently achieves superior performance across all instruction-observation pairs compared to the modality-specific approach. Given that the total volume of training data is identical due to fixed training steps, this enhanced performance is attributed to learning from multiple modalities, which aids in better model generalization.

2. The significance of embodied alignment. The impact of embodied alignment is specifically explored in our experiments presented in Table 2. We experiment by removing the embodied alignment module and simply concatenating the outputs of projection layers from different modalities. TokenLearner is still utilized to manage the memory burden of numerous tokens. The alignment between instruction tokens and observation tokens is then achieved using an MLP layer, followed by the policy network. Notably, there is a significant performance drop when embodied alignment is omitted.

3. Does Incorporating More Modalities Enhance Performance? A key advantage of the Any2Policy framework is its ability to outperform modality-specific models by training on multiple modalities for instruction-observation pairs. It raises the question: can including more modalities at the inference stage further improve performance? Table 4 presents our experimental findings. We used the text-image pair as the baseline, a common setup in robotics. Adding additional instructional

Table 3: Comparison with state-of-the-art robotic models in real-world experiments.

Instruction → Observation	Image → Image	Text → Image	Video → Image	Image + Text → Image
VIMA [59]	-	-	-	49
R3M [109]	42	-	46	-
T5 [110]	-	39	-	-
Any2Policy	**44**$_{+2}$	**51**$_{+12}$	**59**$_{+13}$	**62**$_{+13}$

Table 4: Ablation study on the effect of using different modalities in real-world experiments.

Method	Instruction	Observation	Success Rate
Any2Policy	Text	Image	51
	Text+Audio	Image	52
	Text+Audio+Image End-Goal	Image	60
	Text+Audio+Image End-Goal + Video Demonstration	Image	**62**
Any2Policy	Text	Image	51
	Text	Image+Video	57
	Text	Image+Video+Point Cloud	**66**
Any2Policy	Text	Image	51
	Text+Video Demonstration	Image+Video	63
	Text+Image End-Goal	Image+Point Cloud	**69**

modalities, such as audio, image-based end-goals, and video demonstrations, consistently increased the success rate. Interestingly, although text and audio convey similar instructions, their combination still yielded a slight performance improvement of 1%. For observations, significant performance gains were observed when adding video and point cloud data, integrating temporal and 3D spatial information. Combining multiple instructional and observational modalities also improved performance; for example, incorporating video demonstrations in instructions and video in observations led to a 12% increase in the success rate.

4. *Comparison of Any2Policy with state-of-the-art robot models.* To further assess the effectiveness of Any2Policy representations, we compare them with other robot models, as shown in Table 3. Any2Policy consistently outperforms both T5 [110] and R3M [109] across various modalities, highlighting the advantage of incorporating multiple modalities during training. Furthermore, Any2Policy significantly surpasses VIMA [59], a recent method that uses text goals and object images for task specification. This demonstrates that Any2Policy not only outperforms single-modal counterparts, but is also highly effective in employing multiple modalities during inference.

5. *Other results.* We observe that the performance of image-point cloud and video-point cloud pairings achieves relatively lower scores compared to other modality pairs. However, when we use text or audio as instruction, the performance of point cloud is typically better than image or video due to its accurate position in three-dimensional space. This indicates that there is still room to improve model performance with point cloud inputs. These results highlight the inherent complexity of integrating visual modalities like images and videos with point cloud data, which often requires sophisticated fusion techniques to effectively combine spatial information with traditional 2D features.

4.2 Simulation Evaluation

Dataset. Our experiments are based on the following three datasets in our simulation. These three datasets represent three different instruction-observations modalities. Specifically, Franka Kitchen [92] uses text-image and ManiSkill2 [94] uses text-image and text-{image, point cloud}.

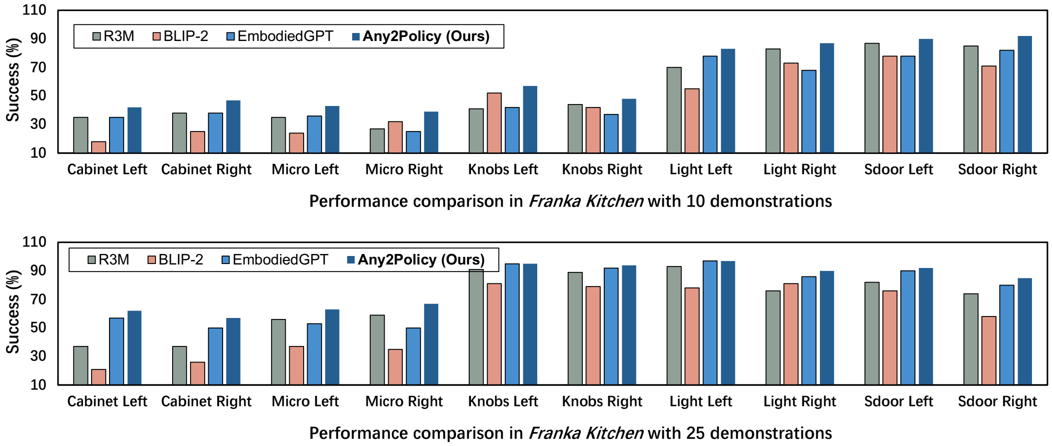

Figure 4: Performance of Any2Policy in Franka Kitchen with 10 or 25 demonstration demos. Comparison with R3M, BLIP-2, and EmbodiedGPT. On all tasks except for Knobs-left with 25 demonstrations, we obtained superior performance over SOTA methods.

The objective of our simulated environments is to evaluate the efficacy of our proposed method under conditions where modalities are limited.

Implementation Details. All models are trained on A100 GPUs, implemented in PyTorch [111]. The Franka-Kitchen are trained for 40K steps. We use weight decay of 1e-6, cosine learning rate scheduler with warmup steps of 2% total steps. The gradient clip of 1.0 is also applied. We use Adam optimizer with initial learning of 1e-3 and 3e-4 for Franka Kitchen and Maniskill-2. Note that Maniskill-2 does not have text instructions. We augment the task description into instruction via GPT-4. It is then tokenized and prepend with image backbone via FiLM [112].

Evaluation: We evaluate our approach using 30 rollouts from the BC learned policy. We use the mean success rate of the final policy as our metric. When providing task suite metrics we average the mean success rate across camera configurations.

Experimental results on Franka Kitchen. In our experiments, we benchmarked our model against two state-of-the-art methods: R3M [109], and BLIP-2 [22], a sota vision-language model, and Embodied-GPT [113], a multi-modal model designed for robotics. Our policy network was trained using few-shot learning, utilizing either ten or twenty-five demonstrations. We assessed the success rate of these models in 100 randomized evaluations across five different tasks in each benchmark. These evaluations were conducted under two settings, each with five separate runs and from two different camera perspectives, using only visual observations. The results, illustrated in Figures 4 for the Franka Kitchen, clearly show that Any2Policy outperforms the baseline methods, on both 10 and 25 demonstrations, except for Knobs-left with 25 demonstrations.

5 Conclusion

Humans interact with and perceive the world through information from multiple modalities. Existing research in robotics predominantly concentrates on policy learning using a single modality. In contrast, this paper introduces the Any2Policy framework, which adeptly demonstrates the integration of various sensory modalities into a cohesive model. This framework efficiently processes and responds to multi-modal data for robotic tasks. The overall framework, coupled with its multi-modal dataset, represents a significant advancement in the field of embodied AI. Our findings underscore the potential of Any2Policy to augment robotic adaptability and performance, thereby highlighting the importance of multi-modal methodologies in the development of more generalized and versatile robotic systems.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[2] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[3] 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.

[4] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe 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.

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[6] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021.

[7] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[8] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin 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.

[9] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.

[10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

[11] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021.

[12] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[13] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019.

[14] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.

https://doi.org/10.52202/079017-4244

[15] Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 34:16064–16078, 2021.

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

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

[18] 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.

[19] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[23] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[24] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

[25] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[26] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

[27] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[28] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[29] Xinke Jiang, Rihong Qiu, Yongxin Xu, Wentao Zhang, Yichen Zhu, Ruizhe Zhang, Yuchen Fang, Xu Chu, Junfeng Zhao, and Yasha Wang. Ragraph: A general retrieval-augmented graph learning framework. *NeurIPS*, 2024.

[30] Xinke Jiang, Zidi Qin, Jiarong Xu, and Xiang Ao. Incomplete graph learning via attribute-structure decoupled variational auto-encoder. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 304–312, 2024.

[31] Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. *arXiv preprint arXiv:2312.15883*, 2024.

[32] Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

[33] 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.

[34] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[35] 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.

[36] Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia Pinel, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pages 175–187. PMLR, 2023.

[37] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

[38] 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.

[39] Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation. *arXiv preprint arXiv:2409.14411*, 2024.

[40] 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.

[41] Yixuan Wang, Zhuoran Li, Mingtong Zhang, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D ˆ 3 fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. *arXiv preprint arXiv:2309.16118*, 2023.

[42] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[43] Shizhe Chen, Ricardo Garcia, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. *arXiv preprint arXiv:2309.15596*, 2023.

[44] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[45] Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on Robot Learning*, pages 284–301. PMLR, 2023.

[46] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*, 2024.

[47] Ge Yan, Yueh-Hua Wu, and Xiaolong Wang. Dnact: Diffusion guided multi-task 3d policy learning. *arXiv preprint arXiv:2403.04115*, 2024.

https://doi.org/10.52202/079017-4244

[48] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros, and Trevor Darrell. Zero-shot visual imitation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2050–2053, 2018.

[49] Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint arXiv:2210.10047*, 2022.

[51] 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.

[52] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[53] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. First person action-object detection with egonet. *arXiv preprint arXiv:1603.04908*, 2016.

[54] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[55] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[56] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[57] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019.

[58] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019.

[59] 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. *arXiv preprint arXiv:2210.03094*, 2022.

[60] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multimodal task specifications. *arXiv preprint arXiv:2309.14320*, 2023.

[61] Yichen Zhu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Retrieval-augmented embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17985–17995, 2024.

[62] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022.

[63] Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. *arXiv preprint arXiv:2212.05749*, 2022.

[64] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*, pages 17359–17371. PMLR, 2022.

https://doi.org/10.52202/079017-4244

[65] Kaylee Burns, Tianhe Yu, Chelsea Finn, and Karol Hausman. Pre-training for manipulation: The case for shape biased vision transformers.

[66] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[67] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[68] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[69] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, et al. Bridgedata v2: A dataset for robot learning at scale. *arXiv preprint arXiv:2308.12952*, 2023.

[70] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[71] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[72] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[73] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[74] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[75] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-ϕ: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024.

[76] Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. A comprehensive overhaul of multimodal assistant with small language models. *arXiv preprint arXiv:2403.06199*, 2024.

[77] X Liu, Y Zhu, J Gu, Y Lan, C Yang, and Y Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *arXiv preprint arXiv:2311.17600*, 4(1):4–4, 2023.

[78] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[79] Felix Hill, Sona Mokra, Nathaniel Wong, and Tim Harley. Human instruction-following with deep reinforcement learning via transfer-learning from text. *arXiv preprint arXiv:2005.09382*, 2020.

[80] 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.

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

[82] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[83] Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. Mmro: Are multimodal llms eligible as the brain for in-home robotics? *arXiv preprint arXiv:2406.19693*, 2024.

[84] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024.

[85] Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. *Advances in neural information processing systems*, 30, 2017.

[86] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020.

[87] Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Zhengping Che, Chaomin Shen, Yaxin Peng, Dong Liu, Feifei Feng, and Jian Tang. Language-conditioned robotic manipulation with fast and slow thinking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4333–4339, 2024.

[88] Junjie Wen, Yichen Zhu, Minjie Zhu, Jinming Li, Zhiyuan Xu, Zhengping Che, Chaomin Shen, Yaxin Peng, Dong Liu, Feifei Feng, and Jian Tang. Object-centric instruction augmentation for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4318–4325, 2024.

[89] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[90] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *7th Annual Conference on Robot Learning*, 2023.

[91] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[92] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[93] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[94] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

[95] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10809–10818, 2022.

[96] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 784–794, 2022.

[97] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[98] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2092–2101, 2023.

[99] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *Advances in neural information processing systems*, 34:11960–11973, 2021.

[100] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

[101] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[102] Qiqi Zhou and Yichen Zhu. Make a long image short: Adaptive token length for vision transformers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 69–85. Springer, 2023.

[103] Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. *Advances in Neural Information Processing Systems*, 35:32011–32024, 2022.

[104] Yichen Zhu, Qiqi Zhou, Ning Liu, Zhiyuan Xu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Scalekd: Distilling scale-aware knowledge in small object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19723–19733, 2023.

[105] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021.

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

[107] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[108] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

[109] 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.

[110] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[111] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[112] 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.

[113] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction concisely outline our framework for retrieving information from video to enhance visuomotor learning. Our experiments support that our framework improves the generalization of the robot model.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

https://doi.org/10.52202/079017-4244

Justification: Our method contains strong assumptions that affordance in human video is similar to manipulation. Therefore, our proposed method may not benefit non-human like embodiment such as locomotion learning.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We have not theoretical result.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the necessary information to reproduce our method, including the training details, the dataset we used, the implementations, and the setup of our environment for the real robot.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justfication: The data will be attached in the webpage.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

https://doi.org/10.52202/079017-4244

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Because the evaluation of robot manipulation is extremely time-consuming, we report the average success rate over 20 trails on our real-world experiments. For the simulation environment, we follow previous methods which report the average success rate over 30 trails. Details are specified in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computer resources.

https://doi.org/10.52202/079017-4244

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We follow the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We specify the potential positive/negative societal impacts in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

https://doi.org/10.52202/079017-4244

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the assets and codes that we used in our work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

https://doi.org/10.52202/079017-4244

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-4244