Beyond Sight: Finetuning Generalist Robot Policies with Heterogeneous Sensors via Language Grounding

Joshua Jones*, Oier Mees*, Carmelo Sferrazza*, Kyle Stachowicz, Pieter Abbeel, Sergey Levine

https://fuse-model.github.io

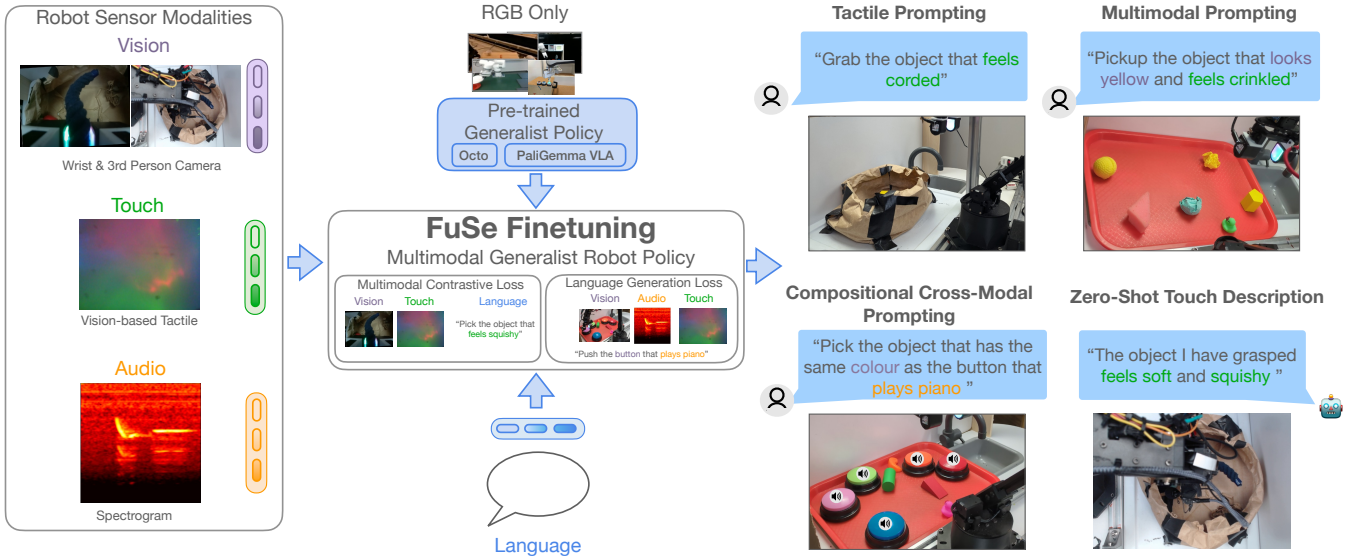

Fig. 1: We introduce FuSe, an approach that enables finetuning large image-based pre-trained generalist policies, including vision-language-action (VLA) models, on heterogeneous robot sensor modalities, such as touch or audio, for which large datasets are not readily available, while leveraging natural language as a common cross-modal grounding. Our finetuning recipe enables challenging multimodal and cross-modal prompting tasks in partially-observable scenes and is able to generate zero-shot descriptions of objects it interacts with.

***Abstract*— Interacting with the world is a multi-sensory experience: achieving effective general-purpose interaction requires making use of all available modalities – including vision, touch, and audio – to fill in gaps from partial observation. For example, when vision is occluded reaching into a bag, a robot should rely on its senses of touch and sound. However, state-of-the-art generalist robot policies are typically trained on large datasets to predict robot actions solely from visual and proprioceptive observations. In this work, we propose FuSe, a novel approach that enables finetuning visuomotor generalist policies on heterogeneous sensor modalities for which large datasets are not readily available by leveraging natural language as a common cross-modal grounding. We combine a multimodal contrastive loss with a sensory-grounded language generation loss to encode high-level semantics. In the context of robot manipulation, we show that FuSe enables performing challenging tasks that require reasoning jointly over modalities such as vision, touch, and sound in a zero-shot setting, such as multimodal prompting, compositional cross-modal prompting, and descriptions of objects it interacts with. We show that the same recipe is applicable to widely different generalist policies, including both diffusion-based generalist policies and large vision-language-action (VLA) models. Extensive experiments in the real world show that FuSe is able to increase success rates by over 20% compared to all considered baselines.**

*Equal contribution, authors listed alphabetically. The authors are members of Berkeley AI Research (BAIR), UC Berkeley, USA. Please email correspondence to the lead authors: {joshuajones, csferrazza, oier.mees}@berkeley.edu.

I. INTRODUCTION

Intelligent beings have the ability to seamlessly combine a variety of sensory feedback that allows them to effectively interact with physical the world. Beyond vision, humans rely on the touch and audio feedback to manipulate objects [1], [2], as they provide rich complementary information about object properties, especially when visual information alone might be insufficient to complete the task, such as when locating keys inside a bag [3]. This stands in contrast to state-of-the-art robot policies [4]–[8], often denoted as *generalist*, that absorb knowledge from a vast amount of robotics datasets [9]–[13], but rely solely on visual and proprioceptive observations to perform a wide range of tasks.

The main factor limiting development of generalist robot policies based on truly hetereogeneous data is that, while recent robotics datasets contain an abundance of vision and proprioception data, only a small minority of them contain other sources of sensory data [14]–[16]. This raises the question: how can we retain the generalization capabilities of generalist robot policies pre-trained on large amounts of data, while connecting their semantic knowledge with heterogeneous sensory data, for which large datasets are not readily available?

Prior studies show that natural language can act as common grounding across different models, even when they are

trained on minimally overlapping data domains [17]–[22]. Moreover, relating human language to multimodal percepts and actions naturally enables indexing goals using open-vocabulary multimodal queries. Nonetheless, incorporating multiple sensing modalities, such as touch or audio, into robotic policies has thus far proved challenging. This difficulty arises from factors such as data scarcity, the tendency of prior work to focus on single-sensor approaches, and the lack of joint reasoning over multimodal percepts and low-level robotic actions [2], [3], [14]–[16], [23]–[26].

In this work, we address these challenges and present a recipe to finetune generalist robot policies on smaller-scale datasets comprising modalities complementary to vision, such as touch and sound, and demonstrate that novel capabilities and cross-modal semantic understanding are unlocked through this multimodal finetuning procedure.

Our key insight is to use language as a bridge across all modalities. By doing so, we enable our policy to perform challenging manipulation tasks that require reasoning jointly over vision, touch, and sound in a zero-shot setting, enabling multimodal prompting, generation of object descriptions upon interaction, and compositional cross-modal prompting. In practice, our policy can successfully fulfill challenging task instructions, such as "pick the red object that feels soft and makes a loud sound", "describe how the grasped object feels like", "pick the object that has the same color as the button that plays piano".

Our results show that our policies leveraging a pre-trained generalist robot policy finetuned on multimodal data consistently outperform baselines finetuned only on vision data, or trained from scratch on heterogeneous sensory data. We find that the same general recipe is applicable to generalist policies with widely different architectures, such as Octo [4], a large transformer-based policy trained on the Open X-Embodiment [9] (OXE) dataset, and a 3B VLA with a PaliGemma [27] vision-language-model VLM backbone.

For our experiments, we leverage a dataset consisting of 27K robot trajectories we collected that contains vision, touch, audio, proprioception, and language instructions on three different real-world robotic manipulation tasks. To the best of our knowledge, this dataset is the first of its kind that also contains robot action data, which is key to perform physically grounded multimodal tasks. We open-source all of our data, code and models to support future research in this area.

II. RELATED WORK

A. Generalist Robot Policies

Generalist robot policies have shown promise of consuming diverse large-scale data to unlock generalization in robotic tasks [4]–[8], [22], [28]. These policies leverage large robot dataset collections [9], [10], [29] that have recently been made available to the community, and are most often queried with language instructions defining the task. In some instances, robot actions are directly fused with a vision-language model (VLM) backbone [5], [7], [22], improving generalization due to pre-training on internet-scale data.

However, while some of the recently introduced models [4], [8] can naturally process flexible observations, the scarcity of datasets that include other sensory modalities, such as touch or audio, limits their capabilities primarily to visual inputs. In contrast, our work shows how such capabilities can be enhanced with a much smaller amount of robotic data containing additional heterogeneous modalities to allow jointly reasoning over modalities such as vision, touch, and sound in a zero-shot setting.

B. Multimodal Reasoning in Robotics

Multimodality aims to exploit complementarity across different sensors to enhance the capabilities of autonomous robot policies. Its advantages have repeatedly been shown in the literature, resulting either in improved performance [2], [3], [3], [24], [30]–[39], generalization [31], [40], or robustness [37], [41].

Despite this evidence, only a minority of works employ sensor modalities in addition to vision and proprioception. This is reflected in the robotics datasets made available to the community. For example, the largest collection of robotics dataset, Open X-Embodiment [9] (OXE), does not include touch or sound as part of their default sensory modalities. Some notable exceptions include recent works [14], [23], [42] that try to align vision, language, and touch for perception tasks. However, most of the available datasets made available through these works do not include robot actions, limiting their applicability for policy training and to perform physically grounded multimodal tasks. Here, we first introduce a multi-task dataset that includes vision, touch, audio, inertial measurements, proprioception, as well as robot actions and language instructions. We then leverage this dataset to finetune large generalist robot models, unlocking novel multimodal reasoning capabilities.

III. FUSE FINETUNING

State-of-the-art generalist robot policies typically rely on vision, language, and robot actions as training modalities, which limits their applicability on partially-observable scenes where tasks cannot be completed solely through vision. We propose a recipe, FuSe, to **Fu**se heterogeneous **Se**nsory data into generalist robot policies. Specifically, we finetune these policies to extend their semantic understanding to include additional sensing modalities, such as touch and sound, while retaining their pre-trained knowledge. Our key observation is that by adding two auxiliary losses, which contrast heterogeneous observations with natural language and generate language from observations, we are able to link a variety of sensing modalities with the semantic knowledge of pre-trained generalist robot policies. We use Octo [4], a transformer-based pre-trained policy, as the backbone model for the main experiments in this paper, but we also show that the same finetuning recipe is applicable to a 3B vision-language-action model based on a PaliGemma [27] VLM backbone. The training architecture is depicted in Figure 2.

This finetuning strategy introduces three main challenges, namely: (i) the weights of the feature extractors (encoders)

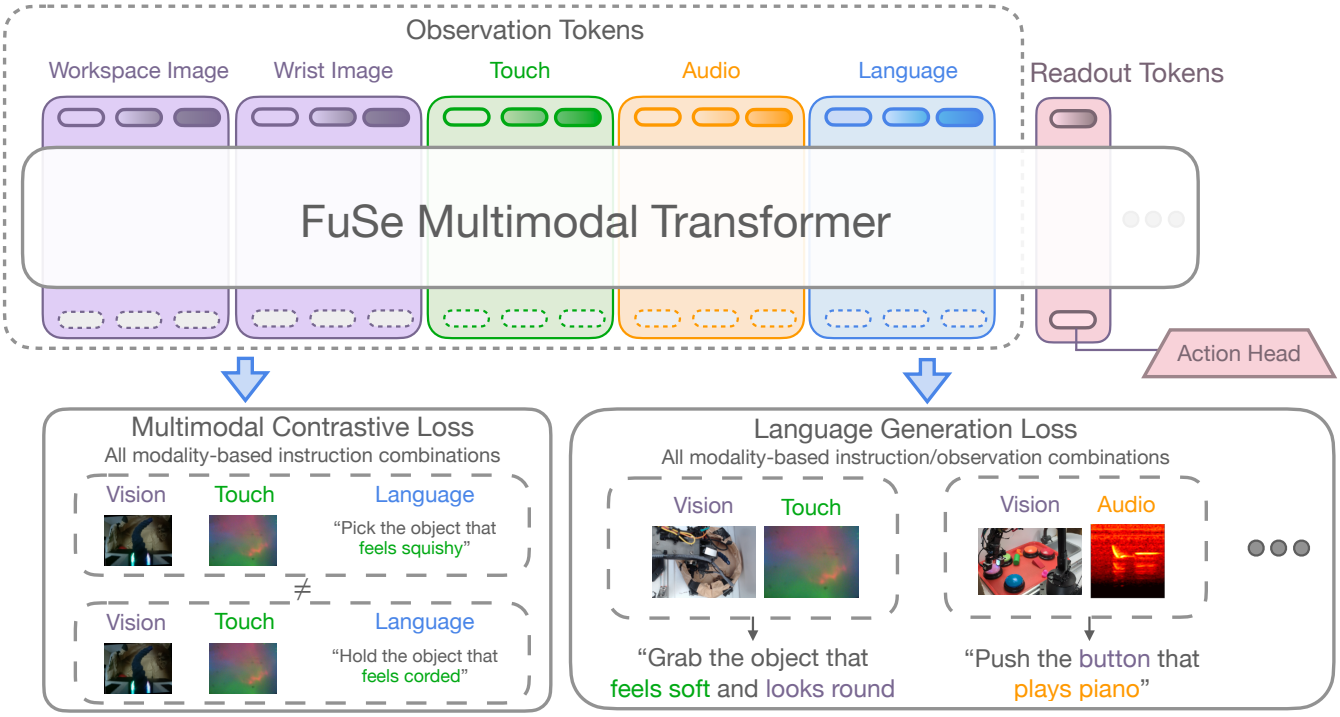

Fig. 2: **Architecture:** We finetune pre-trained generalist robot policies by tokenizing all heteregoneous sensing modalities and passing them though a pre-trained transformer backbone. Crucially, we apply two auxiliary losses that help connect the semantic knowledge of pre-trained generalist policies with new heterogeneous modalities, such as touch and audio. Concretely, we apply both a contrastive loss that aims to maximize mutual information between different views and semantics of the same scene, and a language generation loss that predicts high-level semantics for each modality combination.

for the new modalities generally need to be effectively learned from a small dataset; (ii) the finetuned model empirically tends to predominantly rely on the pre-training modalities, ignoring the new sensors; (iii) novel cross-modal prompting capabilities rely on modality specific annotations, e.g., "the object feels soft and squishy". We detail below the modifications required to address all of these challenges.

Tactile encoder. To account for the small finetuning dataset size, we use a pre-trained tactile encoder and finetune it together with the backbone Octo architecture. In particular, we use the TVL encoder [14], which was pre-trained via pairwise contrastive learning across vision, language, and tactile modalities. We feed all tactile images (two in our robot setup) separately through the same TVL encoder.

Audio encoder. As the raw audio waveform is highly dimensional and noisy, we process the audio data to build a spectrogram as reported in previous work [3], [43]–[45]. The spectrogram is then treated as a regular image and fed through a ResNet26 encoder [46].

Auxiliary losses. As aforementioned, a naïve way of simply finetuning pre-trained generalist policies with a mean-square-error (MSE) imitation loss $\mathcal{L}_{BC}$ conditioned on additional sensor data, leads to the policy over-relying on its pretraining modalities and ignoring the new modalities. We overcome this limitation by introducing two additional losses that fully leverage multimodality and connect the semantic knowledge of pre-trained generalist policies with unseen sensor modalities:

1) *Multimodal Contrastive Loss:* We introduce a loss that aims to align the various language instructions with the observations via CLIP-style contrastive learning [47]. At a high level, it aims to maximize mutual information between different modalities and semantics of the same scene. Concretely, we build an observation embedding by feeding all modalities once more through the transformer and combining them via a multi-head attention layer. We then compute a CLIP loss for each possible instruction resulting from combining the different available modalities. These losses are finally averaged to form a combined multimodal contrastive loss $\mathcal{L}_{contrast}$.
2) *Multimodal Generative Loss:* We design a generative network that functions as an add-on head to the backbone model. In practice, for each possible modality combination, we build an observation embedding as above, and feed it through the generative head. Then, we compute an auxiliary cross-entropy loss $\mathcal{L}_{gen}$ by comparing the head output with the appropriate language instruction. We use a single transformer as the generative head for all possible modality combinations, with modality tokens to distinguish between input modalities.

The final loss is given by $\mathcal{L} = \mathcal{L}_{BC} + \beta\mathcal{L}_{gen} + \lambda\mathcal{L}_{contrast}$, where the contrastive loss and the generative loss are summed to the MSE action loss during training.

Language Rephrasing. As discussed previously, cross-modal prompting capabilities require modality specific annotations, e.g., "the object feels squishy and looks round". We annotate the robot trajectories we collect with heterogeneous sensors with after-the-fact language annotations. We annotate

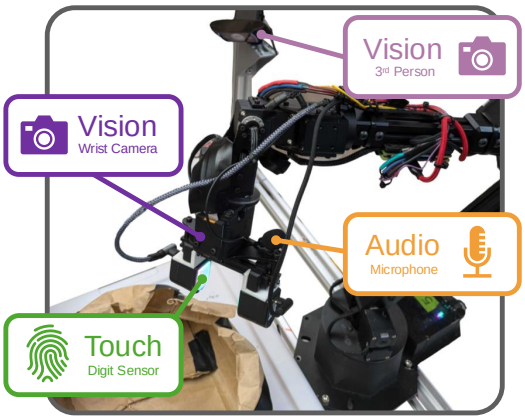

Fig. 3: Visualization of the various sensor modalities on our WidowX robot.

these trajectories with templated language that enables us to create augmentations based on multiple sensor inputs, "the object feels squishy and is red" or "the object feels metallic and sounds clinking". However, at test time we would like users to instruct the policy with free-form language. Therefore, to increase the range of possible input instructions, as well as the representation power of the generative network, we augment the instructions in the dataset by diversifying them through a large language model. Specifically, we query ChatGPT [48] for rephrased templates that preserve the original semantic meaning.

Implementation Details. We train all models for 50,000 steps on a v5e-128 TPU pod with a batch size of 1024. We use a cosine learning rate scheduler with 2000 warmup steps, and a peak value of 3×10^{-4}. We resize third-person RGB images to 256x256, wrist RGB images to 128x128, and tactile images to 224x224. We create spectrograms of size 128x128. Our language rephrasing buffer contains 20 different templates for each possible modality combination. We set $\beta = 1$ and $\lambda = 1$.

IV. Experiments

In this section, we investigate the effectiveness of FuSe to finetune pre-trained generalist robot policies to include additional sensor modalities, while linking them to the policy's pre-trained semantic knowledge. We answer the following questions:

1) **Does FuSe help perform multimodal prompting tasks in a zero-shot manner in partially observable environments?** (Section IV-C)
2) **Does FuSe enable multimodal prompting to discriminate between objects that would otherwise be ambiguously described through a single modality?** (Section IV-D)
3) **Can the multimodal capabilities of FuSe be exploited for compositional reasoning tasks?** (Section IV-E)
4) **Are the proposed cross-modal language grounding losses necessary to achieve high performance when finetuning FuSe?** (Section IV-F)
5) **Is FuSe applicable to different generalist robot policies?** (Section IV-G)

(a) Objects used for evaluation purposes. (b) Objects included in the training data.

Fig. 4: Visualization of objects for real-world experiments, including objects seen (a) and unseen (b) in the multimodal finetuning dataset. Objects differ in shape, appearance, material, hardness, and surface properties.

A. Real Robot Setup and Training Data

All our experiments feature a WidowX 250 6-DoF robot arm. The robot is controlled via delta end-effector position commands at a frequency of 5 Hz. The system is equipped with a third-person view RGB camera, a wrist RGB camera, two DIGIT tactile sensors at the gripper fingers, a standard microphone, and a 9-DoF IMU. We present experiments on three different tasks, which are described below. For the grasping scenarios, we evaluate on the 24 objects present in the training dataset, along with 32 unseen test objects; for the button tasks, we evaluate on the six buttons and 13 of the 18 distractors/grasping targets seen in the training dataset, as well as two unseen buttons and 12 unseen distractors. We visualize the training and test objects used in Figure 4.

We evaluate each model on several different scenarios (e.g., different objects and distractors) for each of the tasks, by running the same scenario for 5 different rollouts.

We collect a dataset of 26,866 trajectories, where the robot is teleoperated using a Meta Oculus Quest 2 VR headset. Each trajectory is labeled with a templated language instruction. The two grasping tasks (tabletop and shopping bag) feature visual, tactile, and action data, while the button pressing tasks also includes sound. Visual observations are recorded at a resolution of 640x480, while DIGIT images at a resolution of 320x240. We follow previous work and perform background subtraction on the tactile images to further emphasize the membrane deformation and reduce systematic differences across DIGIT instances [2]. The audio observations comprise 1s of the most recent microphone samples, recorded at a frequency of 44,100Hz. We visualize our robot sensory setup in Figure 3.

B. Evaluation Tasks

We design a challenging suite of tasks, which focuses on testing the policies' ability to reason jointly over vision, sound, and touch in a zero-shot setting.

Tabletop Grasping. We set up a simple tabletop grasping scenario, where multiple objects are placed on a tray and the task is to grasp the right object as prompted via a text instruction (e.g., pick the carrot).

Shopping Bag. This environment presents a more complex grasping scenario, where objects are placed inside a paper bag. This scenario generally features occlusions to third-person view camera, as well as results in poor lighting conditions for the wrist camera as soon as the gripper enters

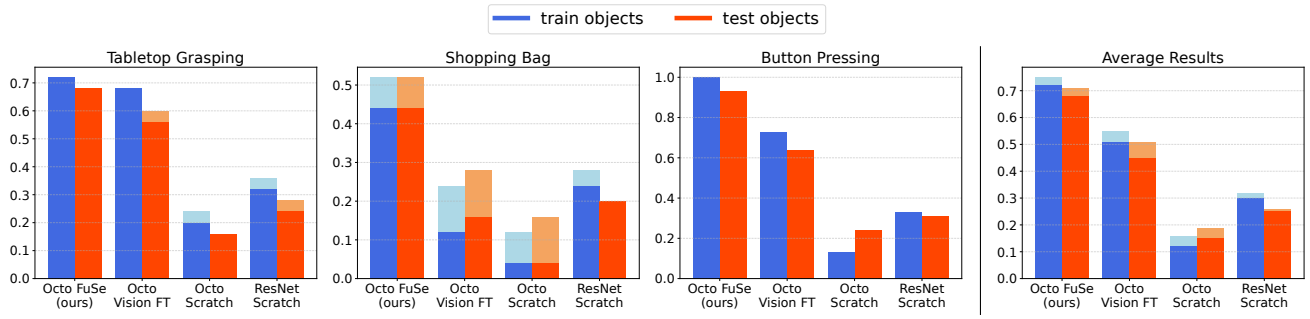

Fig. 5: FuSe performance on evaluation tasks compared against baselines. Our approach outperforms baselines trained from scratch or finetuned with vision only, especially on the shopping bag task, which presents partially observable visual scenarios. Lighter shades of color represent intermediate task success, i.e., object touched but not fully grasped.

the bag. Thus, this represents an environment with partially-observable visual scenarios.

Button Pressing. In this environment, we leverage the sound modality, featuring six sound-making buttons, each playing different sounds upon pressure. The goal is to press the right button depending on the prompt, which can present either visual- or audio-related commands (e.g., "press the red button", "press the button that plays piano", etc.). We also devise two compositional tasks in this setting, where the objective is either i) to grasp objects that share visual characteristics with one of the buttons (e.g., "grab the object that has the same color as the button that plays piano"), or ii) to press among the training buttons the one that plays the same sound as an unseen button (e.g., "press the button that plays the same sound as the blue button").

C. Finetuning Performance

Here, we investigate the benefits of finetuning the Octo generalist policy, pre-trained on the large OXE robotics dataset [9], on our multimodal dataset. First, we are interested on whether our model performs better than the same architecture trained from scratch on a small dataset as ours. The results in Figure 5 show how our approach largely surpasses training Octo from scratch on our multimodal dataset without our finetuning recipe, which is challenging due to the limited size of the dataset. In contrast, our approach leverages the knowledge acquired during pretraining and can adapt to the new tasks and modalities with a smaller amount of additional data. Finally, we also compare against a ResNet26 baseline, where language instructions are fed through FiLM conditioning [49] as done in [50]. The smaller ResNet26 performs slightly better than training Octo from scratch, but still underperforms our model on all three tasks.

To validate the effect of the new modalities on finetuning performance, we compare with a recipe that finetunes Octo only using the available pre-trained modalities, i.e., vision and action. The results in Figure 5 show how this baseline is competitive on the simpler tasks (tabletop and button pressing), but it is considerably inferior to our model on the bag task, where visual occlusions make visual features less discriminative when the gripper enters the shopping bag.

D. Multimodal Prompting

In addition to improving finetuning performance, our training recipe provides the model with additional multimodal capabilities, such as the possibility to provide a multimodal prompt that can successfully discriminate objects based not only on visual features but also based on other modalities such as touch or sound. The evaluation prompts contain several instances where the task is to grab an object with an ambiguous description for one modality, but unique for another (e.g., "grab the round object that feels squishy", where the scene presents both a foam ball and a crumpled paper ball). The results are shown in Table I for the grasping tasks, on scenarios that present objects sharing the same visual and tactile features, respectively. This experiment demonstrates that our policy can incorporate multimodal instructions to improve over ambiguous descriptions.

E. Compositional Capabilities

Finally, we showcase compositional capabilities of our model with two different compositional tasks in the button pressing environment:

- In a simpler task, we prompt the model to grab an object that has the same color as the training button the plays a certain sound (e.g., "grab the object with the same color as the button that plays piano").
- In a multi-step task, we exploit the generative head to connect between different subtasks. First, we prompt the model to press a button not seen at training time, using only visual instructions (e.g., "press the blue button"). Then, we feed the resulting sound to the generative head, which will generate the instruction related to the corresponding audio (e.g., "press the button that plays

	Visual		Visual, Tactile	
	Reach	Grasp	Reach	Grasp
Tabletop	0.43	0.43	0.5	0.43
Bag	0.3	0.25	0.55	0.3
Average	0.37	0.34	0.53	0.37

(a) Vision-ambiguous objects

	Tactile		Visual, Tactile	
	Reach	Grasp	Reach	Grasp
Tabletop	0.4	0.4	0.4	0.4
Bag	0.35	0.3	0.5	0.3
Average	0.38	0.35	0.45	0.35

(b) Touch-ambiguous objects

TABLE I: Multimodal prompting results obtained with the FuSe policy on scenarios that present objects sharing the same visual (a) or tactile (b) features. Our policy incorporate multimodal instructions and improves over ambiguous descriptions.

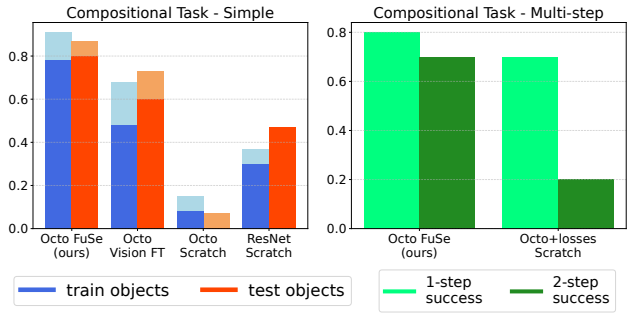

Fig. 6: Results on the compositional tasks devised in the button pressing environment. On the left, the instructions are of the type "pick the *object* that has the same color as the button that play *piano*". On the right, the whole multi-step task is represented by an instruction like "press the train button that plays the same sound as the *blue* button".

piano"). Finally, we prompt the model with the audio instruction in the training environment, where the model has already associated the visual cues of the button to the corresponding sound, and will execute a trajectory that ends up pressing the button that plays the same sound as the button pressed in the first subtask.

We report quantitative results in Figure 6, showing that even on the simple compositional task, FuSe outperforms all baselines, exploiting its multimodal reasoning capabilities. For the multi-step task, we compare with Octo trained from scratch on all available sensors and with the same auxiliary losses. Once again, FuSe outperforms the baseline, particularly on the full task completion. In fact, the model trained from scratch shows poor language grounding and does not succeed in fulfilling the audio-based instruction.

F. Auxiliary Losses Ablation

In this section, we ablate the different FuSe auxiliary losses in the shopping bag task, which features partially observable visual scenarios. Figure 7 shows that both losses are key to fully exploit the heterogeneous feedback available on the robot, with the performance particularly deteriorating for the baselines on test objects.

G. Vision-Language-Action Model Results

We also investigate the effectiveness of FuSe to finetune alternative generalist policies based on off-the-shelf *vision-language-action* (VLA) models. Instead of Octo, we finetune

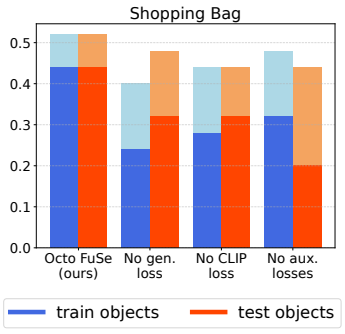

Fig. 7: We study the effect of the proposed losses in an ablation experiment in the shopping bag environment. Our model that includes both contrastive and language generative losses outperforms models trained with only one of the two auxiliary losses or neither.

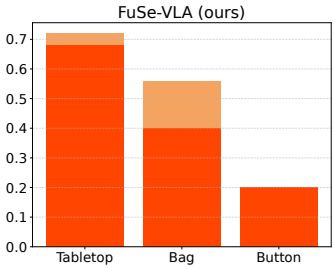

Fig. 8: Performance of a PaliGemma FuSe 3B parameter VLA, trained on our multimodal dataset, on unseen test objects. Our policy achieves robust performance on the grasping tasks, showcasing the applicability of FuSe to widely different generalist policies.

a 3B parameter vision-language model to get FuSe-VLA, a VLA model capable of producing both robot actions and language grounding. We use the PaliGemma [27] VLM as the backbone, as it is able to easily incorporate a flexible set of observations modalities (similar to Octo, but unlike other VLA models like OpenVLA [5]). Such models are also able to incorporate FuSe's generative language modeling loss directly rather than requiring an additional language model head, unifying the implementation of action prediction and language-based feature learning. We first pre-train on the Bridge dataset [11], and finetune on our dataset with all sensor modalities. We show results for the FuSe-VLA on unseen test objects in Figure 8. These preliminary results demonstrate that FuSe shows promise to transfer across different robot policy architectures. We note that the difference in performance on the button pressing task may be ascribed to the Bridge dataset being only a subset of OXE, which instead contains button pressing tasks among its trajectories.

To our knowledge, FuSe-VLA is the first open-source VLA finetuned on heterogeneous sensory inputs.

V. CONCLUSIONS

In this paper, we introduced FuSe, an approach to finetune large, pre-trained robot policies on heterogeneous robot sensor modalities, such as touch or audio, for which large datasets are not readily available. By leveraging natural language as a common cross-modal grounding during training, FuSe enables performing challenging tasks that require reasoning jointly over modalities such as vision, touch, and sound in a zero-shot setting. FuSe enables capabilities such as multimodal prompting, compositional cross-modal prompting, and descriptions of objects it interacts with. We also demonstrate the effectiveness of our recipe (multimodal finetuning and feature learning via cross-modal language grounding) is applicable to widely different generalist policies, including a transformer-based Octo model or a policy finetuned from a generative VLM base model pre-trained on internet-scale data as well as unimodal robot data.

A limitation of our approach is that training a policy with additional modalities requires increasing training resources, which currently limits our observation history to 0.4s. Increasing training efficiency would enable training with longer context length, potentially leading to improved reasoning about sparse signals such as tactile data, and will be subject of future work.

REFERENCES

[1] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, 2009.

[2] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[3] M. Du, O. Y. Lee, S. Nair, and C. Finn, "Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning," *arXiv preprint arXiv:2205.14850*, 2022.

[4] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

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

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[8] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," in *Conference on Robot Learning*, 2024.

[9] O. X.-E. Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, H. Su, H.-S. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Kim, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Wu, J. Luo, J. Gu, J. Tan, J. Oh, J. Malik, J. Tompson, J. Yang, J. J. Lim, J. Silvério, J. Han, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Zhang, K. Majd, K. Rana, K. Srinivasan, L. Y. Chen, L. Pinto, L. Tan, L. Ott, L. Lee, M. Tomizuka, M. Du, M. Ahn, M. Zhang, M. Ding, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, P. R. Sanketi, P. Wohlhart, P. Xu, P. Sermanet, P. Sundaresan, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Moore, S. Bahl, S. Dass, S. Song, S. Xu, S. Haldar, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Dasari, S. Belkhale, T. Osa, T. Harada, T. Matsushima, T. Xiao, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Li, Y. Lu, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. hua Wu, Y. Tang, Y. Zhu, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Xu, and Z. J. Cui, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024.

[10] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[11] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[12] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard, "Latent plans for task agnostic offline reinforcement learning," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, Auckland, New Zealand, 2022.

[13] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[14] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Goldberg, "A touch, vision, and language dataset for multimodal alignment," in *Forty-first International Conference on Machine Learning*.

[15] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "Maniwav: Learning robot manipulation from in-the-wild audio-visual data," in *8th Annual Conference on Robot Learning*.

[16] N. Cheng, Y. Li, J. Gao, B. Fang, J. Xu, and W. Han, "Towards comprehensive multimodal perception: Introducing the touch-language-vision dataset," *arXiv preprint arXiv:2403.09813*, 2024.

[17] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[18] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

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

[20] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[21] ——, "Audio visual language maps for robot navigation," in *Proceedings of the International Symposium on Experimental Robotics (ISER)*, Chiang Mai, Thailand, 2023.

[22] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[23] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens *et al.*, "Binding touch to everything: Learning unified multimodal tactile representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 340–26 353.

[24] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *Conference on Robot Learning*. PMLR, 2017, pp. 314–323.

[25] T. Bi, C. Sferrazza, and R. D'Andrea, "Zero-shot sim-to-real transfer of tactile control policies for aggressive swing-up manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5761–5768, 2021.

[26] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5633–5640.

[27] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[28] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiullah, "Robot utility models: General policies for zero-shot deployment in new environments," *arXiv preprint arXiv:2409.05865*, 2024.

[29] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[30] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, "General in-hand object rotation with vision and touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 2549–2564.

[31] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," *arXiv preprint arXiv:2311.00924*, 2023.

[32] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016.

[33] J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch:

Audio-visual pretraining for contact-rich manipulation," *arXiv preprint arXiv:2405.08576*, 2024.

[34] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," *arXiv preprint arXiv:2212.03858*, 2022.

[35] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang, "Robot synesthesia: In-hand manipulation with visuotactile sensing," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6558–6565.

[36] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[37] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[38] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto, "See to touch: Learning tactile dexterity through visual incentives," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 825–13 832.

[39] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[40] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.

[41] P. Miller and P. Leibowitz, "Integration of vision, force and tactile sensing for grasping," *Int. J. Intell. Mach*, vol. 4, pp. 129–149, 1999.

[42] S. Yu, K. Lin, A. Xiao, J. Duan, and H. Soh, "Octopi: Object property reasoning with large tactile-language models," *arXiv preprint arXiv:2405.02794*, 2024.

[43] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[44] ——, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[45] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[46] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[48] J. Schulman, B. Zoph, C. Kim, J. Hilton, J. Menick, J. Weng, J. F. C. Uribe, L. Fedus, L. Metz, M. Pokorny *et al.*, "Chatgpt: Optimizing language models for dialogue," *OpenAI blog*, vol. 2, no. 4, 2022.

[49] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

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