See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation (Supplementary Materials)

Hao Li[1*] **Yizhi Zhang**[1*] **Junzhe Zhu**[1] **Shaoxiong Wang**[2] **Michelle A Lee**[1]
Huazhe Xu[1] **Edward Adelson**[2] **Li Fei-Fei**[1] **Ruohan Gao**[1†] **Jiajun Wu**[1†]
[1]Stanford University [2]Massachusetts Institute of Technology
[*]Equal contribution. [†]Equal advising.

The supplementary materials consist of:

1. Supplementary video.
2. Generalizing to new initial conditions for the pouring task.
3. More qualitative results.
4. Details of the task setup.
5. Details of our multsensory self-attention model.
6. Ablation study on using the Franka robot's force sensing.
7. Ablation study on using the wrist-mounted camera.
8. Ablation study on the roles of magnitude vs. frequency in audio

1 Supplementary Video

In the supplementary video, we show 1) the motivation of leveraging multisensory data for robotic manipulation tasks and our multisensory learning pipeline, 2) illustrations of the dense packing task and qualitative results, and 3) illustrations of the pouring task and qualitative results.

2 Generalizing to New Initial Conditions for the Pouring Task

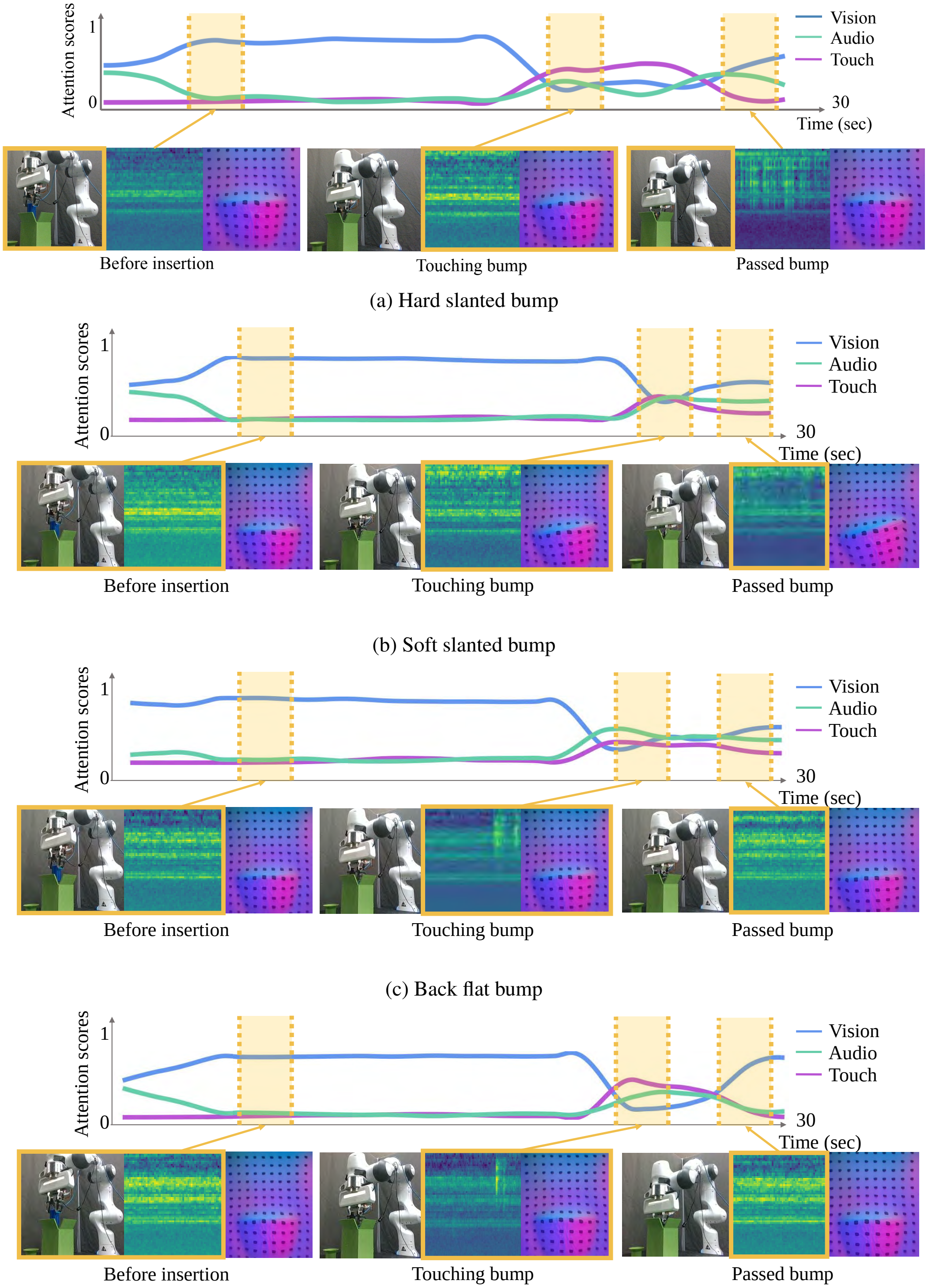
For the pouring task, we have conducted two generalization tests with initial weight of $110g$ and $120g$. Table 1 summarizes the results. We show the mean weight error of the beads poured into the fixed cup as we display in the main paper. We can see that our model perform wells in both circumstances, even if the initial conditions are out of the training distribution.

With the increased initial weight, it is more difficult for the model to capture the dynamics of the beads, leading to larger average error. Our MULSA model still achieves the best results, outperforming the two baseline methods that use a different mechanism for fusing the three modalities.

Initial weight (g)	110	120	Average
Direct Concat	3.27 ± 1.34	19.8 ± 1.85	11.54 ± 1.54
Du et al. [1]	4.91 ± 3.59	15.33 ± 2.19	10.12 ± 3.72
MULSA (ours)	**1.51 ± 0.83**	**3.82 ± 1.16**	**2.66 ± 1.02**

Table 1: Mean weight error of beads poured into the cup (mean ± standard deviation) (g).

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.



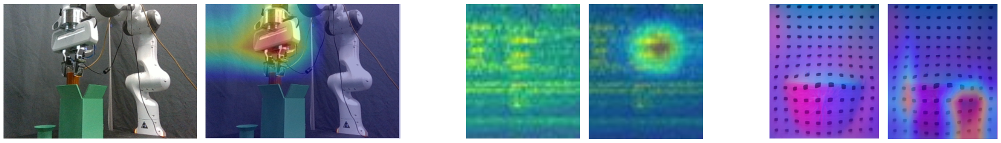
(a) Hard slanted bump

(b) Soft slanted bump

(c) Back flat bump

(d) Left flat bump

Figure 1: Aggregated attention scores on all four bases of the dense packing task: (a) hard slanted bump, (b) soft slanted bump, (c) back flat bump, and (d) left flat bump. In all four scenarios, we demonstrate the observations at three key moments, when the robot aligns the peg with the base before insertion (left), the peg touches the different bumps (middle), and the peg has already passed the bump and can be pushed down into the base (right).



(a) Vision (b) Audio (c) Touch

Figure 2: Examples of the heatmaps showing the focused area within each observation of (a) vision, (b) audio, and (c) touch for some key moments during test trials. The image for vision is captured when the peg is in the process of get aligned with the base, and the sensory observations for audio and touch are captured at moments of contact.

3 More Qualitative Results

3.1 Visualization of Attention Scores

In Fig. 5 of the main paper, we have visualized the aggregated attention scores for each modality as the robot completes the dense packing task with the hard slanted base and the pouring task. In Fig. 1, we show the results for the test trials of the other three settings of the dense packing task, which are computed in the same way as the hard slanted case. Specifically, the multisensory attention layer takes $N = 6$ encoded observations from each of the three modalities, yielding a 3N×3N raw attention score matrix. Then at each time step, we sum the attention scores on all N embeddings of each modality to obtain the three aggregrated attention scores.

It can be seen that when testing on all four bumps, the attention is mostly on vision before insertion starts or after the peg has already passed the bump. This is because during these time periods, the model either needs to rely on vision to align the peg with the base, or can simply push the peg down to complete the task with no need to rely on audio or touch.

Most failures of the packing task happen because the peg touches the bump but then takes the wrong action, therefore the model must attend to the correct information to distinguish the local geometry. In all four cases, the attention shifts towards audio and touch when contact occurs because the visual inputs are similar. For the two slanted bumps, we see from Fig. 1a that the hard surface generates a high frequency acoustic signal while the soft surface (Fig. 1b) does not, indicating that the peg is scratching on the bump. If the surface is hard, the correct policy is to avoid the bump, therefore the attention stays on the touch for a longer time in the hard bump case than the soft one. In the two flat bump scenarios (Fig. 1c and Fig. 1d), we observe very different acoustic signals than in the slanted bump cases, which suggest that the contact mode has now switched to collision. The model then attends to touch to check which direction the peg is tilting toward, so that it can infer the location of the bump and how to avoid it.

3.2 Grad-cam Visualization

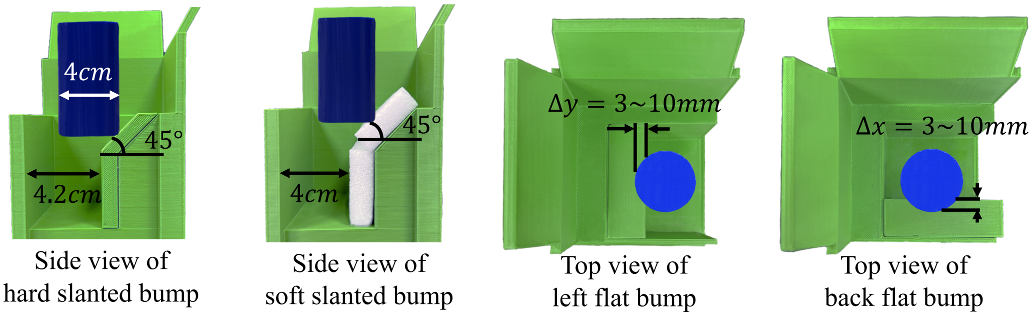
To visualize where the model is attending to within each observation, we use Grad-cam[2] to generate gradient-based localization heatmap obtained from our MULSA model tested on the packing task as shown in Fig. 2. We can see that for vision, it focuses on the end effector position to align the peg with the base (Fig. 2a). The focus within the audio segment is roughly on the sound frequencies caused by the contact, as shown in Fig. 2b. Lastly, for touch, it focuses on the edge of the peg contour as seen in Fig. 2c, as most dynamic changes of the peg can be detected from these areas.

4 Details of the Task Setup

In Sec. 3.3 of the main paper, we introduced the goal and setup for the two robotic manipulation tasks that we tackle: dense packing and pouring. In this section, we provide more details on the setup of these two tasks.

4.1 Dense packing

Setup details. In the dense packing task, we use four types of bases with different geometry designs to test the effectiveness of our multisensory model and the roles of different sensory modalities for this task. The task is to fit the in-hand object, which in our experiment is a cylinder peg, into the empty space left in the base. The peg has a diameter of 4cm and a height of 6cm. Inside the base,



Side view of hard slanted bump

Side view of soft slanted bump

Top view of left flat bump

Top view of back flat bump

Figure 3: Illustration of the four types of the bases we use for the dense packing task.

we design four different bumps to create more diverse contact modes, with details shown in Fig. 3. Two of the bumps use an L-shaped bump with a 45° slope followed by a vertical wall. This slanted bump is located in the front of the base, leaving a 2mm tolerence at the end of the slope. These two bumps share the same geometry, yet their surface materials differ. We leave one base with the hard 3D-printed surface, and pad the other one with a layer of soft sponge. The other two bases use hard flat bumps without any padding, with one having the bump located on the left side and the other one having the bump located on the back. During both training and testing, the position of the flat bumps randomly varies in a $7mm$ range.

In the real-world dense packing task, the goal is to fit a glass at the bottom of a crowded box without destroying other objects around it. The box is densely packed with various objects and there is only one hollow space left in the middle, as shown in Fig. 1 of the main paper. In this setup, we have different objects including a thin metal board, a soft stuffed toy, a towel, and a plastic plate. During the process, the glass will interact with these objects, and the robot must respond differently to avoid the obstacles. With the camera, the robot is able to locate where the empty spot is, yet there are small objects arranged in the bottom of the box which cannot be visually observed, therefore the robot also needs audio and touch to correctly identify the geometry and material of the objects.

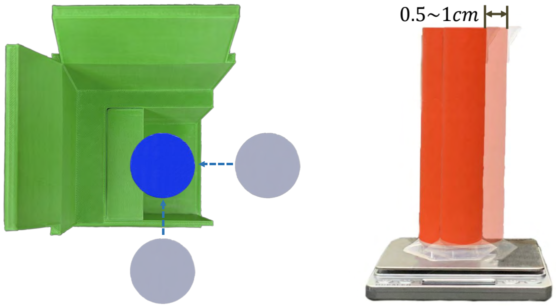
Robot action space. In this task, the end effector can move either horizontally to adjust the position, or vertically to insert the object. The human expert gives command from a set of actions in x, y, z dimensions, with each action step size being $\delta x = 0.3mm, \delta y = 0.3mm, \delta z = 1mm$. In each dimension, the action could be $\pm\delta$ or 0, thus the size of the action space is $3^3 = 27$. The same action space applies to the model prediction.

Human policy and data collection. In each trial, the peg is first randomly initialized to the right or back of the base, as illustrated in Fig. 4a. We need to move the peg horizontally to align it with the base, then we can start to push it downward. After the peg makes contact with the base, each of the four bases corresponds to a different human expert policy. On the hard slanted bump, the correct policy is to move backward ($-\delta x$) to avoid the bump, then continue inserting ($-\delta z$) along the back wall. If the surface is soft, the policy is to keep pushing and squeeze the peg down into the slot, which makes the packing sufficiently dense and also increases efficiency. When the peg touches either flat bump, the robot must move in the opposite direction to avoid the bump.

When collecting human demonstrations, we limit the observation space of the human experts to the same level as the robot sensors by only showing them the real-time multisensory data. The operator cannot perceive the entire environment during the experiment, and they must decide the correct next step for the robot by monitoring the data captured by our visual, acoustic, and tactile sensors.

4.2 Pouring

In the pouring task, we have the robot pour small beads from one cup into another up to a specific level, resembling human pouring water. We use the weight of the beads as our quantitative metric, and the goal is to pour 40g beads. Testing is performed under a slightly different setting outside the training setting distribution, where the fixed cup is shifted by 0.5∼1cm, as illustrated in Fig. 4b.



(a) Dense packing

(b) Pouring

Figure 4: Random initialization during testing in the (a) dense packing, and the (b) pouring tasks.

Setup details. We use 1mm diameter beads made of procelain for pouring. A 150ml cylinder container is grasped by the robot, and another 250ml container is fixed on the table to catch the dropping beads. Initially, the in-hand container either has a weight of 60g or 100g.

Robot action space. In this task, the robot can act in a 2-dimensional action space along dimensions x, ϕ, where x corresponds to the horizontal movement and ϕ is the rotation. The action step size is $\delta x = 0.3mm, \delta\phi = 0.07°$. The action in each dimension could be $\pm\delta$ or 0, leading to $3^2 = 9$ action space size.

Human policy and data collection. In the human demonstration, the correct policy in a trial can be divided into three stages. In the first stage, the end-effector keeps rotating in $-\delta\phi$ direction, and moves forward $(+\delta x)$ to align the two containers. Once the beads start dropping, the robot stops rotating to reduce the flow rate for more precise control. In the second stage, the action remains 0 in both x and ϕ. In the final stage, the robot stops pouring and rotates in the $+\phi$ direction to retrieve the container.

In this task, human operators perform action by eyeballing the locations of the containers to align them, and decide when to stop pouring by monitoring the electronic weight scale.

5 Model Details

Our model takes 18 observations (6 frames of visual and tactile observations, and 3 seconds of acoustic signal divided into 6 equally long segments), and encodes each of them using a ResNet-18 [3]. For vision and touch, we perform an adaptive average pooling on the final feature map, then flatten and downsample the feature vector to the dimension of 128 using one fully-connected layer. For audio, the mel spectrogram representation is fist computed from the raw waveform. We also change the channel dimension in the first convolutional layer from 3 to 1 to accommodate the 1D acoustic input, and the rest of the audio encoder is the same as the encoders used for the other two modalities as discussed above. All 18 embeddings we obtain from the three modalities share the same embedding size of 128. These embeddings are passed into the joint self-attention layer, which uses a multi-headed self-attention with 8 heads. The attention layer is followed by a 3-layer MLP, with the first two layers being fully-connect with 1024 layer size, and the last layer projects the hidden embedding to logits over the task-dependent discrete actions and has a Softmax after it.

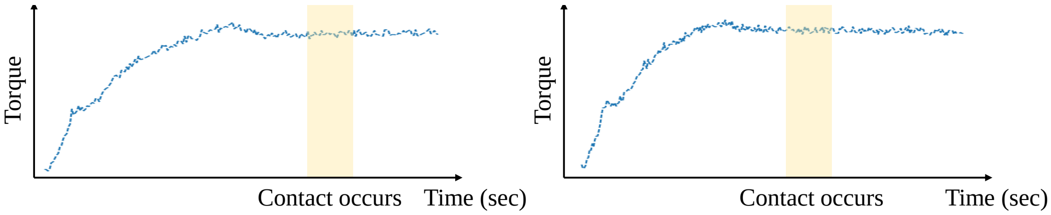
6 Ablation Study on Using the Franka Robot's Force Sensing:

One commonly used type of robot sensor is the force-torque sensor, which records the force and torque exerted by each of the robot joints. To compare the model performance using the force-torque sensor with our current results using vision, audio, and touch, we demonstrate an ablation study on the dense packing task.

In this experiment, we substitute the audio and tactile sensors with the Franka arm's force-torque (FT) sensor. At each time step, the FT sensor returns 7 torque values corresponding to the 7 joints of the robot. We stack 6 groups of FT readings that span 3 seconds of time history together, yielding

	Hard slanted	Soft slanted	Back flat	Left flat	Avg. succ. rate
Table-mounted + Force-torque	0.00	**1.00**	0.00	0.00	0.25
Table-mounted + Wrist-mounted	**1.00**	0.00	**1.00**	**1.00**	0.75
Table-mounted + A + T (Ours)	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 2: Ablation study results of using the Franka robot's force sensing and the wrist-mounted camera for the dense packing task.



(a) Soft slanted bump

(b) Hard slanted bump

Figure 5: Torque of the end effector joint during the dense packing task on the (a) soft slanted bump and the (b) hard slanted bump. The moment of the peg making contact with the base is labeled in each subfigure.

6×7 values as the force-torque observation. This observation is encoded with a similar method used in [4]. The FT encoder network is composed of four 1D convolutional layers to cast the input vector into a 128 dimensional feature vector. The dimensions of the four convolutional layers are 16, 32, 64, 64, respectively. The downstream network shares the same structure as introduced before, where the encoded FT and visual embeddings are passed into our joint self-attention layer and the model finally outputs an action prediction.

The experiment results are shown in Table 2. The robot can only succeed on the soft slanted case, where the robot does not need to take action to avoid any obstacles and can just push the object straight down. The results show that audio and touch provide more helpful information than the force-torque sensor to the robot. The main reason behind this is that the audio and tactile sensors are highly sensitive to local dynamics, and any change leads to an immediate change in the corresponding signals. However, the force-torque sensor only shows a change when the contact force gets stronger. As shown in Fig. 5, the force-torque sensor readings from the last joint of the robot do not demonstrate any noticeable changes when the contact happens, therefore the robot cannot use the feedback as an evidence to distinguish between various scenarios. However, the contact moment can be well captured with audio and touch, so the model is able to reach a higher success rate compared to using the force-torque sensor.

7 Ablation Study on Using the Wrist-Mounted Camera:

To capture more detailed local information of the environment, one alternative approach is to use a wrist-mounted camera to capture a closer view of the surroundings of the in-hand object. We demonstrate a set of ablation study experiments to evaluate the benefit of using a camera mounted to the end-effector of the robot, as shown in Fig. 6. The wrist-mounted camera is also a RealSense D435 camera that streams with the same setting as the table-mounted camra.

We perform experiments using the original table-mounted camera with the wrist-mounted camera to compensate for the detailed local geometrical information. The results are shown in in Table 2. From the results, we observe that the wrist-mounted camera can help the robot perform well on the flat bumps. However, there is a noticeable drop in the success rate on the slanted bumps compared to the results obtained using audio and touch. Although the wrist-mounted camra suffers less from occlusion so that it can reveal some visible characteristics about the geometry, such as the location of the bump, it is less sensitive to implicit clues related to object dynamics, such as the sound and object deformation during the contact. These invisible features captured in the acoustic and tactile signals help the robot make important decisions about the object characteristics that cannot be easily perceived by many other sensors, therefore it outperforms the other baselines.

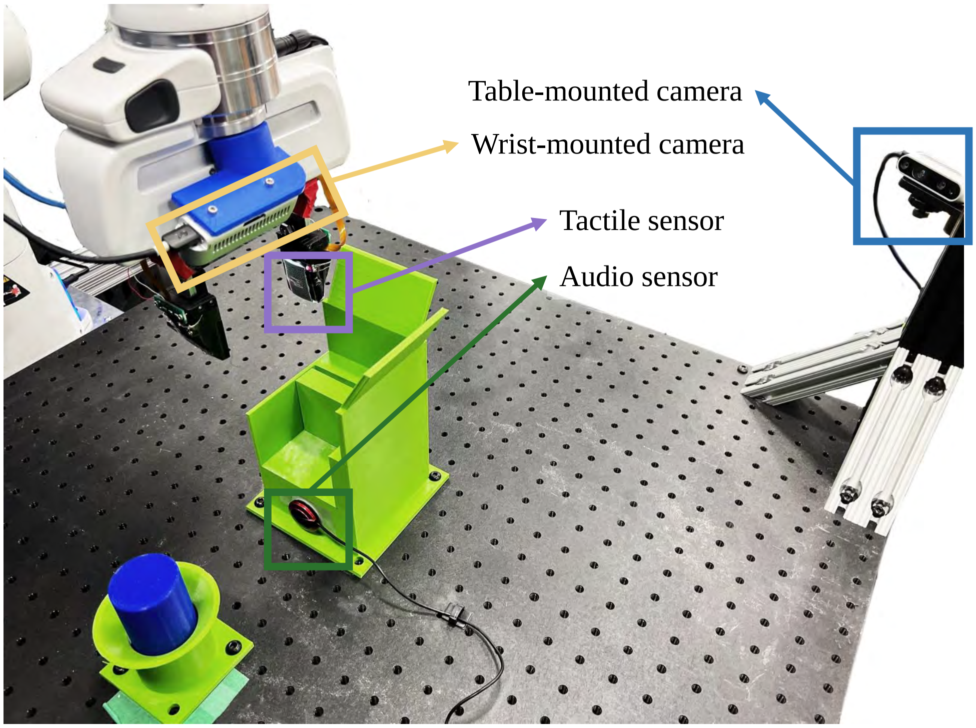


Figure 6: Setup including the wrist-mounted camera.

Initial beads weight ()	60
MULSA (V+A(frequency only)+T)	7.50±1.46
MULSA (V+A(magnitude only)+T)	19.58±0.71
MULSA (ours, V+A+T)	**1.06±0.83**

Table 3: Error of beads poured out (mean ± standard deviation, unit:).

8 Ablation Study on the Roles of Magnitude vs. Frequency in Audio

To further examine the useful information in the acoustic signal, we perform an additional study on the pouring task where we only use either the magnitude or the frequency cues of the recorded audio. Specifically, to remove the information of audio magnitude, we normalize the magnitude in each time window and only keep the relative magnitude scale among different frequencies. To only keep the information in frequency, we first sum up the magnitude at all frequencies in the same time window, and then assign this value to all frequencies in this window to produce a signal that always has equally large magnitude at each frequency bin.

The results are shown in Table 3. We see that when we only keep the magnitude of the audio, the robot always pours all beads into the cup and does not stop when the desired amount has already been achieved. When the frequency information is present, the robot can predict the amount of beads in the cup more accurately. These observations suggest that frequency provides more information to the robot than the magnitude on this task, which aligns with our intuition. When we pour water into a cup, the pitch gets higher as the water level rises. The sound magnitude also helps the robot because as more beads accumulate in the cup, the vibration caused by new beads dropping down gets damped.

References

[1] M. Du, O. Lee, S. Nair, and C. Finn. Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning. In *RSS*, 2022.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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