TEMU - Tactile Encoder for Manipulation: R3. Analysis and Initial Modifications

Bhaswanth Ayapilla * **Aayush Fadia*** **Megan Lee*** **Parth Singh*** **Ranai Srivastav***
{bayapill, afadia, meganlee, parthsin, ranais}@andrew.cmu.edu

1 Introduction

Manipulation is a ubiquitous challenge in robotics, requiring interaction with objects that vary widely in shape, size, texture, and dynamic properties. Current manipulation policies rely heavily on vision-based modeling to estimate these properties, but vision alone fails under heavy occlusion or when properties such as weight are not directly observable. Tactile sensors provide richer information, but the field lacks the decades of research that have gone into extracting meaningful representations from visual data.

To address this, we train a multimodal model that predicts confidence scores for grasp stability using visual, tactile, joint force-torque, joint trajectory, and gripper force readings over a variable-length window. Building on R1 and R2, this report analyzes the outputs of previously discussed approaches, motivates key modifications, and examines failure cases.

2 Analysis

2.1 Intrinsic Metrics

Intrinsic Metric 1: Modality Confidence Score To diagnose whether modalities are being equally exploited during joint training, we track per-modality confidence scores throughout training, motivated by On-the-Fly Gradient Modulation Peng et al. (2022). This metric directly measures how much each modality contributes to the learning objective, without requiring any modification to the training procedure.

For each modality $u \in \{\text{tac}, \text{rgb}, \text{prop}\}$, we attach a lightweight linear probe to the mean-pooled modality embedding, producing a scalar logit $\hat{y}_i^u$. The confidence score is the probability assigned to the correct class:

$$s_i^u = \begin{cases} \sigma(\hat{y}_i^u) & y_i = 1 \\ 1 - \sigma(\hat{y}_i^u) & y_i = 0 \end{cases} \tag{1}$$

To summarize imbalance over a mini-batch $\mathcal{B}$, we compute the discrepancy ratio:

$$\rho^u = \frac{\sum_{i \in \mathcal{B}} s_i^u}{\frac{1}{|\mathcal{M}|-1} \sum_{v \neq u} \sum_{i \in \mathcal{B}} s_i^v} \tag{2}$$

where $\mathcal{M}$ is the set of active modalities. $\rho^u > 1$ indicates modality u is dominating; $\rho^u < 1$ indicates under-optimization.

Table 1 shows consistent modality imbalance across configurations. The two-modal Tactile+RGB baseline is nearly balanced, but adding proprioception introduces significant skew. This instability persists and shifts unpredictably as more modalities are added.

Intrinsic Metric 2: Pixel Fidelity Score (PFS) To verify that learned visual representations attend to task-relevant image regions, we evaluate GradCAM saliency maps using the Insertion protocol Petsiuk et al. (2018). Let π be the pixel ordering induced by GradCAM importance (descending).

[*] Everyone Contributed Equally

			Dev			
Methods	**F1** ↑	**Acc(%)** ↑	**Loss** ↓	$\rho_{\textbf{tac}}$	$\rho_{\textbf{rgb}}$	$\rho_{\textbf{prop}}$
Tactile + RGB (LSTM)	0.731	77.86	0.548	1.007	0.992	—
Tactile + F/T (LSTM)	0.567	40.46	0.878	0.615	—	1.63
Tactile + RGB + F/T (LSTM)	0.556	51.14	0.859	0.838	0.878	1.331
Tactile + RGB + F/T + GF (LSTM)	0.604	51.53	0.857	0.826	0.852	1.383
Tactile + RGB + F/T + GF + G (LSTM)	0.597	44.27	0.888	1.171	1.181	0.7
Tactile + RGB + F/T (GRU)	0.538	38.93	0.861	0.972	0.969	1.061

Table 1: Per-modality discrepancy ratios ρ^u across model configurations. $\rho^u > 1$ indicates dominance; $\rho^u < 1$ indicates under-optimization. Inactive modalities marked —.

Starting from a Gaussian-blurred (information-destroyed) version of the image, we progressively reveal pixels in order π, recording the model's confidence $c_{\text{ins}}(k)$ after restoring the top-$k\%$. If GradCAM is accurate, confidence should climb quickly as we uncover the truly informative regions. We define the Pixel Fidelity Score as

$$\text{PFS} = \text{AUC}\big(\{c_{\text{ins}}(k)\}_{k=0}^{100}\big)\,, \tag{3}$$

where a high PFS indicates that GradCAM's highlighted pixels genuinely drive the model's predictions as confidence rises rapidly as they are revealed. A PFS near 0.5 means the heatmap is essentially pointing at noise. Table 2 reports PFS only for configurations that include both RGB and tactile image modalities.

Methods	**Acc(%)** ↑	**PFS** ↑
Tactile + RGB (LSTM)	74.81	0.749
Tactile + F/T (LSTM)	54.20	0.502
Tactile + RGB + F/T (LSTM)	51.14	0.499
Tactile + RGB + F/T + GF (LSTM)	59.16	0.501
Tactile + RGB + F/T + GF + G (LSTM)	48.47	0.501
Tactile + RGB + F/T (GRU)	52.20	0.502

Table 2: Pixel Fidelity Score across model configurations.

Intrinsic Metric 3: Hide all-but-one modalities Even though we train models on all modalities, we would like to evaluate whether the model is able to predict a result given only a subset of modalities that it was trained with. This metric will tell us which modality the model relies on most for making it's prediction. We perform this ablation by simply putting in the dataset-wide mean values for all the modalities except for the one we wish to evaluate sensitivity with. We evaluate our est-performing model (Model 5 in Table 5). Our results are shown in Table 3. We find that providing all modalities is helpful, and providing only module at a time decreases performance significantly. The margin is not small, the model goes from 86% accuracy with all modalities to the next-best 51% when given only the RGB image. This also tells us that the model relies heavily on the RGB modality, and then on the Force-Torque (FT) modality. This checks out with the results in Table 2.

Setting	**Accuracy**	**Precision**	**Recall**	**Pred 1**	**Pred 0**	**TP**	**FP**	**TN**	**FN**
all-modalities	0.8659	0.7950	0.8660	683	988	543	140	904	84
V-only	0.6391	0.5120	0.8198	1004	667	514	490	554	113
T-only	0.5302	0.4420	0.9601	1362	309	602	760	284	25
FT-only	0.5913	0.4690	0.6746	902	769	423	479	565	204
G-only	0.5380	0.3825	0.3764	617	1054	236	381	663	391
GF-only	0.3752	0.3752	1.0000	1671	0	627	1044	0	0

Table 3: Hide all-but-one modalities results.

Intrinsic Metric 4:

Intrinsic Metric 4: Prediction from one modality hidden Following on from the previous intrinsic metric, we have the model predict by showing everything except one modality. The results are shown in Table 4. Here an interesting pattern emerges — hiding only Tactile information results in the biggest drop in performance! In Table 3, we saw that giving ONLY Tactile did not gain a lot of accuracy. From here we can conclude that tactile information IS very important PROVIDED that other modalities are present also. It is vital in conjunction with other modalities not on it's own, and removing it from the model hurts performance. Force-Torque comes a close second in terms of performance drop when hidden, which agrees with prior results.

Setting	Accuracy	Precision	Recall	Pred 1	Pred 0	TP	FP	TN	FN
all-modalities	0.8659	0.7950	0.8660	683	988	543	140	904	84
hide-V	0.7834	0.6568	0.8852	845	826	555	290	754	72
hide-T	0.7481	0.6958	0.5837	526	1145	366	160	884	261
hide-FT	0.7546	0.6318	0.8293	823	848	520	303	741	107
hide-G	0.8217	0.7150	0.8724	765	906	547	218	826	80
hide-GF	0.8390	0.7472	0.8628	724	947	541	183	861	86

Table 4: Ablation results by hiding one modality at a time.

2.2 Modifications and their Effects

We study four changes relative to the competitive BiLSTM baseline from R2: (i) **five-modal fusion** with per-modality linear projections before the LSTM, (ii) **ablation of deferred resampling (DRS)**, which targets the minority of examples where the pose-phase label disagrees with the shake-phase label, (iii) **better frozen encoders**—OpenCLIP ViT-L/14 for RGB and T3-large for GelSight, in place of ResNet-50, and (iv) **auxiliary losses with on-the-fly gradient modulation (OGM)**, which attaches per-modality classification heads and adaptively scales their gradients to address modality dominance observed via our confidence score analysis.

Unless noted, training uses the random data split, variable-length grasp+pose windows, sensor standardisation for FT/gripper/gripper-force, SGD (momentum 0.9, weight decay 0.01).

Modification 1: Balanced Projections. Starting from the two-modal ResNet baseline (row 1, Table 5), we extended the input to all five modalities: visual (V), tactile (T), force-torque (FT), gripper state (G), and applied gripper force (GF). Naive concatenation (row 2) drops accuracy by 24%, which we attribute to the large dimensional disparity: raw features have dimensions $d_V = d_T = 2048$, $d_{FT} = 12$, $d_G = 6$, $d_{GF} = 1$, giving a total concatenated dimension of $\sum_m d_m =$ 4,109, in which the visual streams overshadow the proprioception.

We address this by applying a linear projection to each modality before concatenation, with target dimensions $p_V = p_T = 512$, $p_{FT} = 128$, $p_G = 64$, $p_{GF} = 32$, yielding a comparatively balanced $\sum_m p_m =$ 1,248-D LSTM input. This recovers accuracy to 76.34% on the full five-modal configuration, enabling the model to use all the information available in the data instead of heavily relying on a couple of modalities. Detailed visualization shown in Figure 1.

Modification 2: Deferred Resampling Ablation. We train an otherwise identical five-modal balanced-projection model with DRS disabled (row 5, Table 5). Accuracy improves by 4.6 points over the DRS variant (80.92% vs 76.34%), while F1 drops slightly (0.752 vs 0.760). DRS forces the model to see equal proportions of stable and unstable grasps in later training, which improves recall on slips but biases the model toward predicting failure on ambiguous inputs. The DRS implementation was inherited from Kanitkar et al. (2022). The loss instability introduced by DRS activation is analyzed further in Section 3.1.

Modification 3: CLIP ViT-L/14 + T3-large Encoders. Our qualitative analysis revealed that the ResNet-50 visual encoder, pre-trained on ImageNet, fails to attend to task-relevant regions under multi-modal supervision. The *Pixel Fidelity Score* highlights shift from the gripper contact zone to the background as additional modalities are added.

	Configuration	Acc. (%) ↑	F1 ↑
ResNet-50 encoders			
1	V+T — R2 baseline	77.64	0.710
2	V+T+FT+G+GF — R2 baseline	53.42	0.561
3	V+T+FT+GF[1]	77.48	0.720
4	V+T+FT+G+GF[1]	76.34	0.760
5	V+T+FT+G+GF[1,2]	**80.92**	0.752
CLIP ViT-L/14 + T3-large encoders			
6	V+T+FT+G+GF, short run[1,2,3]	69.08	0.692
7	V+T[1,2,3]	73.66	0.685
8	V+T+FT+G+GF[1,2,3]	**83.52**	**0.800**

Table 5: Test accuracy and F1 per configuration; [1] Balanced projection, [2] DRS disabled, [3] CLIP ViT-L/14 + T3-large encoders.

This is expected because a network trained to classify everyday objects has no prior for the subtle deformation patterns that distinguish a stable grasp from a slipping one. We therefore replace both vision encoders with models that have task-appropriate priors: OpenCLIP ViT-L/14 Radford et al. (2021), which yields richer scene understanding for the RGB data, and T3-large Zhao et al. (2024), a foundation tactile encoder trained specifically on tactile sensor contact images.

As shown in rows 6–8 of Table 5, the short run underperforms ResNet but with sufficient training the CLIP+T3 combination outperforms the best ResNet result on both accuracy and F1.

Modification 4: Auxiliary Losses & On-the-Fly Gradient Modulation Motivated by the modality imbalance observed in our confidence score analysis (2.1), we apply On-the-Fly Gradient Modulation (OGM) Peng et al. (2022) to encourage more balanced optimization. Since our per-modality probes are side branches disconnected from the main joint loss, we first introduce auxiliary losses to connect them to the backward pass:

$$\mathcal{L} = \mathcal{L}_{\text{joint}} + \lambda \sum_{u \in \mathcal{M}} \mathcal{L}^u \tag{4}$$

where $\mathcal{L}^u$ is a binary cross-entropy loss on the logit of modality probe u and λ is a weighting hyperparameter. This gives each probe a gradient signal, enabling per-modality gradient scaling.

At each iteration, we compute the gradient scaling coefficient:

$$k^u = \begin{cases} 1 - \tanh(\alpha \cdot \rho^u) & \rho^u > 1 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

where α controls modulation strength. After `loss.backward()`, the gradients of each modality probe are scaled by k^u before `optimizer.step()`, slowing down dominant modalities and allowing under-optimized ones to catch up.

Figure 2 and Table 6 show that OGM successfully rebalances modality contributions across all configurations, with discrepancy ratios converging close to 1.0 for all active modalities. The F1 gains are modest, but accuracy improves quite a bit. This suggests that without OGM, the model was leaning heavily on one modality and essentially ignoring the others, and rebalancing forces it to actually use all the information available to it.

Takeaways from Modifications 2.2: Row 1 of Table 5 establishes our two-modal ResNet baseline at 77.64%. Extending naively to five modalities without per-modality projections (row 2); accuracy drops by 24%. Confirming that the LSTM is not learning optimally when the modality dimensions are extremely disproportionate. Balanced projections (rows 3–4) recover this degradation, with the five-modal run reaching respectable test metrics.

Disabling DRS on top of balanced projections (row 5) pushes accuracy to 80.92%, which is the best ResNet result, indicating that deferred resampling sacrifices accuracy, contrary to the findings in

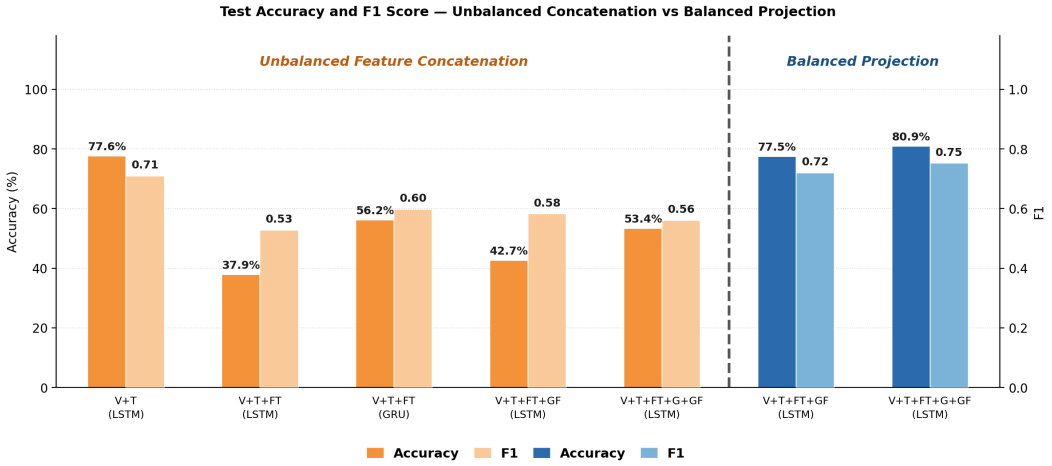


Figure 1: Test Metrics across modality configurations with/without balanced feature projection

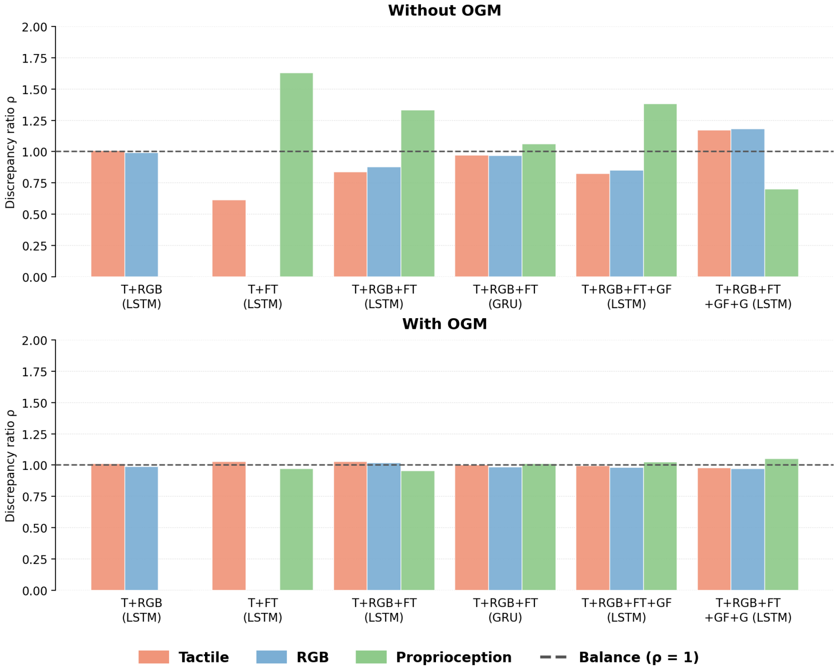


Figure 2: Per modality discrepancy ratios (ρ) before and after On-the-Fly Gradient Modulation

			Dev			
Methods	**F1** ↑	**Acc(%)** ↑	**Loss** ↓	$\rho_{\textbf{tac}}$	$\rho_{\textbf{rgb}}$	$\rho_{\textbf{prop}}$
Tactile + RGB (LSTM)	0.776	79.77	0.424	1.013	0.987	—
Tactile + F/T (LSTM)	0.552	41.22	0.895	1.027	—	0.973
Tactile + RGB + F/T (LSTM)	0.62	60.31	0.879	1.027	1.018	0.956
Tactile + RGB + F/T + GF (LSTM)	0.581	59.16	0.828	0.995	0.981	1.025
Tactile + RGB + F/T + GF + G (LSTM)	0.553	53.43	0.891	0.979	0.972	1.05
Tactile + RGB + F/T (GRU)	0.574	60.31	0.846	1.0006	0.986	1.013

Table 6: Performance and modality discrepancy ratios after applying OGM and disabling DRS

Kanitkar et al. (2022). Among the CLIP+T3 runs, the short run (row 6) underperforms the ResNet baseline, consistent with heavier encoders requiring more iterations to warm up their projection heads; although the two-modal run (row 7) partially recovers. The full-scale run (row 8) achieves 83.52% accuracy and F1 0.800, the strongest result across all configurations, outperforming the best ResNet run by nearly 3%.

Applying OGM (Table 6) successfully rebalances modality contributions, with discrepancy ratios converging close to 1.0 across all active modalities. Accuracy improves consistently relative to the corresponding non-OGM runs, while F1 gains remain modest.

Moving forward to our novel architecture, we adopt no DRS as standard and incorporate OGM to ensure balanced modality optimization, given its consistent improvement in accuracy and convergence of discrepancy ratios across configurations.

2.3 Qualitative Analysis and Examples (full page table)

[2 points]

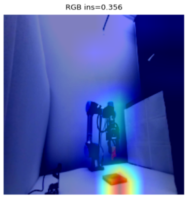
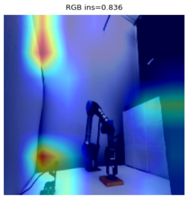
Failure	Example	Description
Visual attention collapse	 Tactile+RGB PFS = 0.749  Tactile+RGB+F/T PFS = 0.499	Adding F/T drops PFS from 0.749 to 0.499($\approx$chance). GradCAM shifts from the gripper contact zone to the background wall, indicating the RGB stream stops attending to task-relevant regions. F/T dominance crowds out the visual gradient signal.
Modality dominance under joint training	Tac+F/T: $\rho_{\text{tac}} = 0.615$, $\rho_{\text{prop}} = 1.63$ Tac+RGB+F/T: $\rho_{\text{tac}} = 0.838$, $\rho_{\text{prop}} = 1.331$	When force-torque is added, proprioception consistently dominates while tactile becomes severely under-optimized. Despite tactile being the most task-relevant modality for slip detection, the model effectively ignores it. The higher-dimensional F/T signal could be capturing more gradient signal early in training and the optimization never recovers.
Accuracy–F1 decoupling after OGM	Tac+F/T before OGM: F1 = 0.567, Acc = 40.46% After OGM: F1 = 0.552, Acc = 41.22%	Even after OGM successfully brings ρ_{tac} and ρ_{prop} close to 1.0, F1 does not meaningfully improve in the Tac+F/T configuration. This suggests that modality imbalance during early training caused the model to learn a suboptimal feature space that cannot be corrected by gradient reweighting alone. The representations are already baked in by the time OGM intervenes.
Predicts only non-slippage when given only Gripper Force Value	Last row in Table 3	This is only one instance of the failure case. It is a result of the imbalance in the dataset, which means it simply learns to predict the most frequent label.
Training instability under deferred resampling (DRS)	Before DRS (iter 0–400): val. acc $\approx 84\%$, loss ≈ 0.44 After DRS activation (iter 400+): loss spikes $>$2.0, val. acc drops to $\approx 70\%$ (Figure 3)	Activating DRS mid-run introduces an abrupt distribution shift: changing the data distribution the model has been optimizing over, causing the observed loss spikes. The resampling strategy is likely not at fault here: a two-stage protocol–pretraining on the full distribution followed by fine-tuning would avoid this instability.

Table 7: Qualitative failure analysis. Each row shows a distinct failure mode observed across model configurations.

3 Impact and insights

Through baselines from previous unimodal experiments and multimodal baselines, it is evident that our current models for the task of grasp-stability prediction suffers from the following 3 challenges:

1. **Training Data Imbalance** - The training data has more samples that exemplify successful grasps, making it difficult for the model to learn semantically meaningful relationships.
2. **Vision Encoders fail on Tactile data** - Most encoders meant for RGB images cannot extract semantically rich information from Tactile sensor outputs.
3. **Performance on unseen objects and poses** - It is easy for the model to learn spurious correlations and not learn semantically meaningful latent representations, leading to overfitting to certain objects and poses.

3.1 Dataset Imbalance

Our dataset is biased towards samples where 66% of the dataset exemplifies experiments that are successful grasps and the rest are failed grasps. To analyze grasp stability, we model the problem as a binary classification problem where we attempt to predict whether a particular grasp will result in a slip or a drop vs neither of the two.

3.1.1 Improving Dataset Imbalance through Deferred Resampling

This dataset imbalance gives the model an incentive to trivially predict successful grasps, leading to an accuracy of around 70%. The model has high accuracy compared to our best trained model thus which has an accuracy of 80% at best on the val set.

To remedy this issue, we subsample our dataset to uniformly distribute successful and unsuccessful grasp samples. This removes the incentive for the model to trivially predict successful grasps. However, this leads to a reduction in the size of the trainable dataset.

To maximize the utility of all the data, we train the model initially on the full dataset to allow the model to learn semantically meaningful latent representations. At a later stage, we initiate subsampling to uniformly distribute the proportion of classes in the dataset.

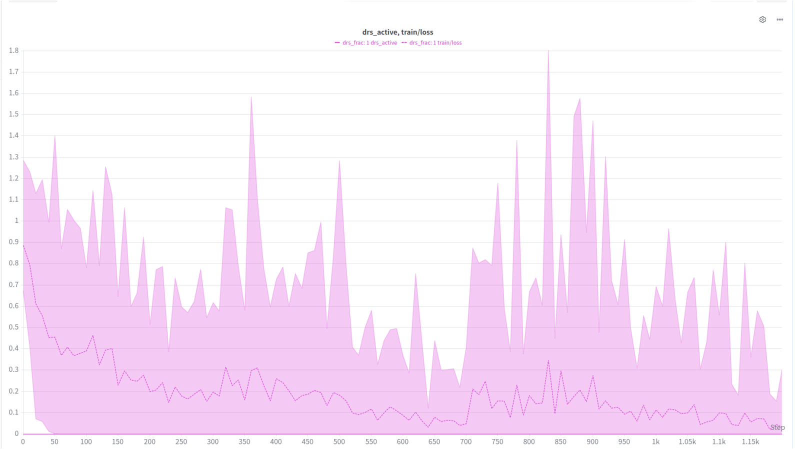


Figure 3: Loss Spikes when DRS is not activated at all. Note the scale of the Y-Axis

Initially, we did naively, during the same training run, leading to a sudden distribution shift leading to training instability. **In future training sessions, we will first conduct a pretraining run to allow the model to learn latent representation for some iterations and then restart training to fine-tune over a subsambled, uniformly distributed dataset.**

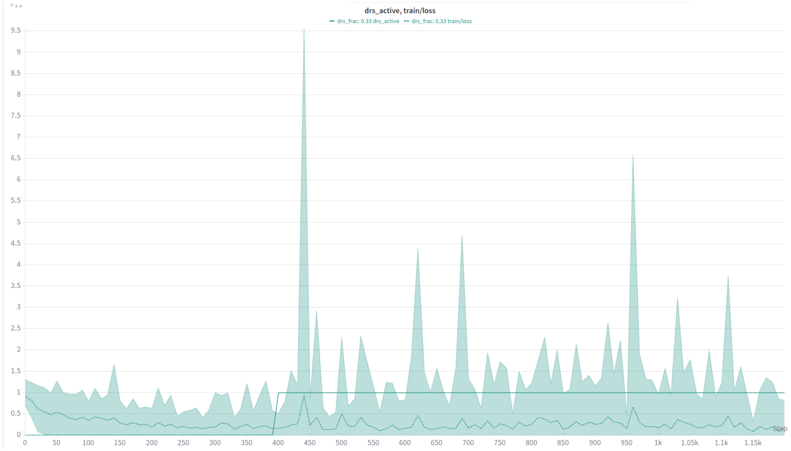


Figure 4: Loss Spikes when DRS is activated after 0.33 of iterations. Note the scale of the Y-Axis

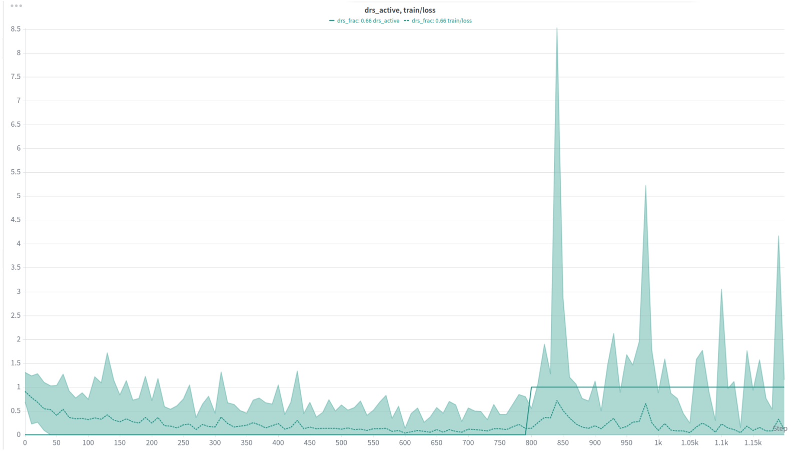


Figure 5: Loss spikes when DRS is activated at 0.66 of total iterations. Note the scale of the Y-Axis

3.1.2 Improving Dataset Imbalance through Tau Normalization

When learning a classification task, the model is trying to learn a decision boundary that divides the space into two distinct subspaces and all input data is projected onto this vector to place it in one of these subspaces. This is represented cannonically as $W^T\mathbf{x}$, which corresponds to our architecture from which predicts one single logit with `BinaryCrossEntropy` loss. Since the dataset has imbalances, the model gets many more updates in the predict success direction and lesser in the other direction, leading to the model being biased towards predicting success. This goes further than spurious correlation because this implies that there is a mathematical bias for the model to predict the majority class.

To mitigate this effect, we convert our architecture from a single logit with `BinaryCrossEntropy` loss to 2 logits with `LogSoftmax`. One logit corresponds to the confidence of predicting one of the 2 classes. This allows us to have 2 decision boundaries $W_{pass}^T\mathbf{x}$ and $W_{fail}^T\mathbf{x}$ normalize the weights of each class with the norm of the weights of that class individually, leading to a unit vector that still represents the normal of the decision boundary and is now not biased because of the size of that vector. As described in Kang et al. (2020), Tau Normalization can reduce the effects of imbalanced datasets, leading to large improvements in long-tailed scenarios.

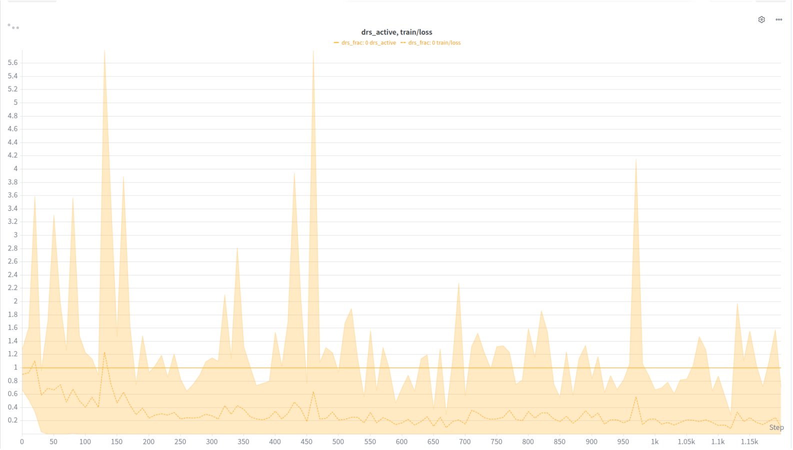


Figure 6: Loss spiked when DRS is activated at the end. Note the scale of the Y Axis

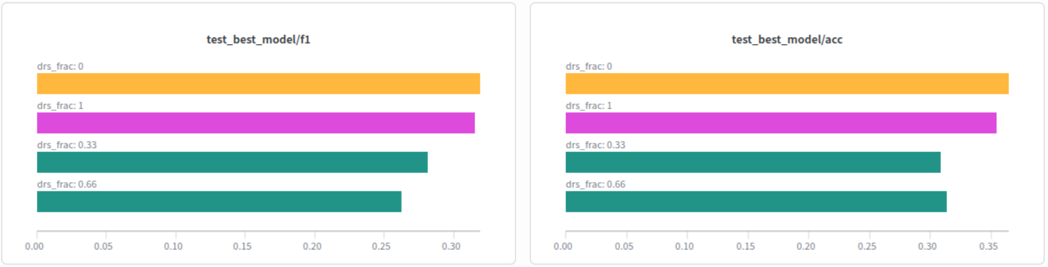


Figure 7: F1 score and Accuracy on the Test Set with and without DRS

3.2 Vision Encoders fail on Tactile data

By the nature of the modality, images are much higher dimensional compared to most other forms of data. 3 channel RGB images of 640 × 480 resolution have little under a million dimensions. Because of this high dimensionality, image modalities require longer training and more data to derive semantically meaningful latent representations. Even though the input is represented in the same way, a vision encoder like ResNet trained on ImageNet will not be able to capture differences that matter in the tactile modality. Good, generalizable encoders, similar to ResNet-ImageNet100 or ClIP, for tactile sensing are an active area for research and currently are hard to come by. Thus it is important that we finetune the vision encoder and do not assume that it is static. When averaged over 10 runs, this accounts for over 10 percentage point difference in F1 score.

To mitigate the effect of this, we plan to pretrain on the full dataset where we finetune the encoder weights for the new CLIP and T-3 configuration.

3.3 Generalization to Unseen Objects and Poses

A persistent risk in grasp stability prediction is that the model learns *object-specific* or *pose-specific* shortcuts rather than contact-driven representations that generalize. For instance, if a cylindrical bottle appears frequently in the training set and is almost always grasped stably, the model may associate its visual appearance with a positive label without ever reasoning about contact forces or tactile deformation.

Our current evaluation uses a random 70/15/15 train/val/test split over the full PoseIt dataset Kanitkar et al. (2022), which does *not* guarantee that test samples come from unseen object identities or novel grasp poses. As a result, the test accuracies reported in Table 5 likely represent an optimistic upper bound on true generalization performance.

Evidence of this risk is already visible in our unimodal ablation (Table 3): the V-only model achieves 63.4% accuracy with zero recall on failures, meaning the visual stream alone has learned to predict "stable" for every input.

Similarly, the PFS scores (≈ 0.499–0.502) across most multi-modal configurations (Table 2) suggest that visual attention is not focused on the contact region, which is the most object-agnostic cue available. Moving forward, a leaving a few objects out for evaluation would provide a more rigorous measure of generalization than the random split used.

The proposed **Multimodal Bottleneck Transformer** in R2 addresses several structural causes of the shortcut learning observed in the LSTM baseline. Because the LSTM receives a single concatenated feature vector at each timestep, the hidden state integrates appearance and contact signals without any mechanism to prevent high-variance modalities from dominating. The bottleneck transformer instead applies self-attention independently per modality before any cross-modal fusion, so each encoder is optimized on its own temporal signal first. Cross-modal fusion is then restricted to a small set of bottleneck tokens that aggregate information from all streams via cross-attention, limiting the degree to which, for example, RGB appearance can propagate into the tactile or force-torque representations.

4 Team member contributions to this report

Bhaswanth Ayapilla contributed to modality confidence score intrinsic metric and OGM modification for the models, also supporting qualitative analysis.

Aayush Fadia Performed the ”hiding” ablations - where one / all-but-one modalities were hidden from our best performing model.

Megan Lee contributed the Pixel Fidelity Score intrinsic metric for the models and report.

Parth Singh contributed the balanced projection, Clip+T3 encoders and DRS ablations.

Ranai Srivastav Worked on model improvements and DRS benchmarking.

References

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020. URL `https://arxiv.org/abs/1910.09217`.

Shubham Kanitkar, Helen Jiang, and Wenzhen Yuan. Poseit: A visual-tactile dataset of holding poses for grasp stability analysis, 2022. URL `https://arxiv.org/abs/2209.05022`.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation, 2022. URL `https://arxiv.org/abs/2203.15332`.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

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*, 2021.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H. Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks, 2024. URL `https://arxiv.org/abs/2406.13640`.