# TEMU - TACTILE ENCODER FOR MANIPULATION: R4. FINAL REPORT

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

## ABSTRACT

As robotic manipulation becomes increasingly ubiquitous in unstructured environments, the ability to predict grasp stability prior to failure is critical for robust execution or failure detection. While integrating complementary modalities—such as vision, tactile, and proprioception—provides a more complete representation of the contact state, naively fusing heterogeneous high- and low-dimensional data streams often degrades performance due to modality dominance. To address this, we propose TEMU, a multimodal bottleneck transformer designed to predict binary grasp stability from 9-15 second temporal sequences of pre-failure sensor data. TEMU extends the Multimodal Bottleneck Transformer (MBT) to five distinct modalities (RGB video, GelSight tactile images, force/torque, gripper state, and commanded force) by restricting cross-modal attention to a compact set of learned bottleneck tokens. By combining this architecture with On-the-Fly Gradient Modulation (OGM) to dynamically balance per-stream optimization, TEMU forces the network to synthesize only the most task-relevant features without allowing high-dimensional streams to overpower weaker signals. Evaluated on the PoseIt dataset, our approach demonstrates that structured representation bottlenecks and balanced gradient updates significantly improve multi-sensory grasp prediction over late-fusion and naive concatenation baselines.

## 1 INTRODUCTION AND PROBLEM DEFINITION

Robust manipulation requires robots to predict whether a grasp will hold before failure occurs. While vision-based policies have driven much of the recent progress in robotic grasping, they struggle when object properties critical to stability (such as weight, surface friction, and compliance) are not directly observable from external cameras. Tactile sensors can supply this missing contact-level information, yet the field lacks the mature pretrained representations and fusion architectures that have powered advances in vision-language modeling. The central challenge, therefore, is not simply *collecting* more modalities but *learning* to combine them: naively concatenating high-dimensional visual features with low-dimensional proprioceptive signals leads to modality dominance, where the model effectively ignores the weaker streams.

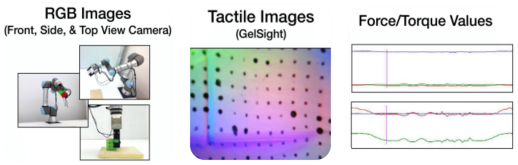



Figure 1: RGB camera views, GelSight tactile contact images, and force/torque readings from Poseit dataset.

We study this problem on the PoseIt dataset (Kanitkar et al., 2022), which provides time-aligned recordings from five sensing modalities: GelSight tactile images, external RGB video, six-axis force/torque readings, gripper state, and commanded gripper force shown in Figure1. It spans 26 everyday objects and up to 16 grasp configurations per object. The task is binary grasp stability prediction: given sensor observations from the *grasping* and *pose* phases of an episode, predict whether the object will be retained or lost during a subsequent shake test. This setting is challenging because the model must anticipate instability from subtle pre-failure cues rather than react to an already occurring slip event.

* Everyone Contributed Equally

Through systematic baselines we show that unimodal models are individually weak (34% accuracy for tactile-only, 61% for RGB-only with near-zero slip recall), and that naive five-modal concatenation *degrades* performance relative to two-modal fusion due to dimensional dominance of the 2048-d visual streams over the 6 to 12-d proprioceptive signals. Balanced per-modality projections recover this loss, and replacing ImageNet-pretrained ResNet-50 backbones with task-appropriate encoders, specifically OpenCLIP ViT-L/14 for RGB and T3-large for tactile, yields the strongest BiLSTM-family result at 83.5% accuracy and 0.800 F1.

To move beyond late-fusion recurrent models, we propose **TEMU**, a multimodal architecture built on the Multimodal Bottleneck Transformer (MBT) (Nagrani et al., 2022). MBT restricts cross-modal information exchange to a small set of learned bottleneck tokens, forcing the network to compress and share only the most task-relevant information across streams. We extend MBT from its original two-modality audio-visual setting to five heterogeneous modalities spanning vision, touch, and proprioception. To address the persistent modality imbalance revealed by our per-modality confidence analysis, we integrate On-the-Fly Gradient Modulation (OGM) (Peng et al., 2022), which adaptively scales per-stream gradients during training to prevent any single modality from dominating optimization.

Our contributions are as follows:

1. A systematic baseline study on PoseIt showing that simple concatenation of modalities results in dimensional dominance, which balanced projections and stronger pretrained encoders largely resolve.
2. An extension of the Multimodal Bottleneck Transformer from two to five modalities, using AdaptFormer adapters and per-stream transformer blocks to handle heterogeneous input dimensions.
3. Two intrinsic diagnostic metrics: a per-modality discrepancy ratio that quantifies optimization imbalance, and a Pixel Fidelity Score that measures whether visual encoders attend to task-relevant contact regions.
4. Ablations over bottleneck size, fusion depth, and modality subsets, revealing that tactile is weak alone but indispensable in fusion, and that gradient reweighting helps but cannot fully undo early modality dominance.

## 2 Related Work and Background

**Related Datasets** Grasp stability prediction has been studied across datasets varying in sensor technology, object diversity, and task formulation. The BiGS dataset (Chebotar et al., 2016) recorded 2,000 grasps with BioTac pressure-array sensors but covered only three objects. Li et al. (Li et al., 2018) scaled to 94 objects using GelSight and an external RGB camera, achieving 88% slip detection accuracy and establishing that fusing vision and touch outperforms either alone, though they omitted force/torque and proprioceptive signals. Calandra et al. (Calandra et al., 2017) further showed that learned tactile features enable grasp success prediction even when vision is ambiguous.

The PoseIt dataset (Kanitkar et al., 2022), which we adopt, addresses many of these gaps with time-aligned recordings from five modalities across 26 objects and up to 16 holding poses (1,840 episodes). Uniquely, PoseIt evaluates stability after the arm repositions and shakes the object, introducing *pose-dependent* stability as the target. The original paper trains a BiLSTM (Hochreiter & Schmidhuber, 1997) and reports 85% accuracy but does not explore transformer-based fusion or modality dominance. Si et al. (Si et al., 2022) tackled a related sim-to-real transfer problem for grasp stability, highlighting how data scarcity and domain shift remain central challenges.

On the representation side, Zhao et al. (Zhao et al., 2024) assembled the Foundation Tactile (FoTa) dataset (3M+ samples, 13 sensors, 11 tasks) to pretrain T3, the first general-purpose tactile encoder. Similarly, Higuera et al. (Higuera et al., 2024) trained the Sparsh self-supervised tactile encoders on 460k+ images and constructed TacBench for evaluation. Both demonstrate that large-scale tactile pretraining substantially outperforms end-to-end training, motivating our use of T3-large.

**Unimodal Baselines** The standard unimodal pipeline passes tactile images through an ImageNet-pretrained ResNet (He et al., 2016) and aggregates temporally via LSTMs (Hochreiter & Schmid-

huber, 1997). Both the PoseIt baseline (Kanitkar et al., 2022) and Li et al. (Li et al., 2018) follow this pattern but note poor transfer to GelSight imagery, since elastomer deformation statistics differ substantially from natural photographs. Dong et al. (Dong et al., 2017) instead detected slip from GelSight using handcrafted surface-motion and shear features, while earlier probabilistic approaches by Bekiroglu et al. (Bekiroglu et al., 2011) showed that even simple classifiers extract useful stability cues when features are well chosen.

For non-visual modalities, Mandil et al. (Mandil et al., 2024) showed that conditioning slip prediction on planned actions improves accuracy, and Nazari et al. (Nazari et al., 2022) coupled a learned slip model with trajectory adaptation for real-time grip correction. These findings motivate our inclusion of gripper state and force as input modalities.

**Prior Work** The PoseIt paper (Kanitkar et al., 2022) fuses modalities by concatenating features per timestep into a BiLSTM, offering no mechanism to prevent 2048-d visual streams from drowning out 6 to 12-d proprioceptive signals. Our experiments confirm that naive five-modal concatenation degrades accuracy by over 24% relative to a two-modal baseline.

The underlying cause is well characterized. Wang et al. (Wang et al., 2020) identified "modality laziness": faster-learning modalities suppress gradient signal to slower ones, locking them into under-optimized representations. Wu et al. (Wu et al., 2022) provided complementary analysis, characterizing the "greedy" convergence where multimodal networks rely on whichever modality provides the easiest shortcut. Peng et al. (Peng et al., 2022) proposed On-the-Fly Gradient Modulation (OGM) as a remedy, adaptively dampening dominant-modality gradients based on a per-stream discrepancy ratio while injecting Gaussian noise to preserve generalization. We adapt OGM to our five-modality setting by attaching per-modality classification probes.

On the architectural side, Nagrani et al. (Nagrani et al., 2022) introduced the Multimodal Bottleneck Transformer (MBT), which restricts cross-modal flow to a small set of learned bottleneck tokens rather than full pairwise self-attention (Vaswani et al., 2017). Each modality is processed independently for the first $L_f$ layers before fusion, achieving state-of-the-art audio-visual classification with roughly 50% fewer FLOPs than a vanilla multimodal transformer. We extend MBT from two modalities to five and average bottleneck updates across all active streams (Eq. 4).

**Relevant Techniques** For RGB encoding we use OpenCLIP ViT-L/14 (Radford et al., 2021), whose vision-language pretraining yields richer scene features than ImageNet-supervised ResNet, and for tactile we adopt T3-large (Zhao et al., 2024). Both visual streams are tokenized with factored spatial-temporal positional embeddings following the ViViT (Arnab et al., 2021) strategy adapted from ViT (Dosovitskiy et al., 2021). Since fully fine-tuning these backbones risks overfitting on PoseIt's 493 episodes, we insert AdaptFormer adapters (Chen et al., 2022) in parallel with each RGB block's FFN, adding less than 2% extra parameters while enabling task-specific adaptation without catastrophic forgetting.

On the training side, PoseIt introduced Deferred Resampling (DRS) (Kanitkar et al., 2022) to address the 65/35 class imbalance by sub-sampling stable grasps after a warm-up phase. However, our analysis reveals that activating DRS mid-training causes sudden distribution shifts producing loss spikes and accuracy drops. Disabling DRS in favor of balanced projections and OGM yields better accuracy (80.9% vs. 76.3%), consistent with Kang et al. (Kang et al., 2020), who found that decoupling representation learning from classifier rebalancing outperforms joint resampling. For visual diagnostics, we use GradCAM (Selvaraju et al., 2017) evaluated with the Insertion protocol (Petsiuk et al., 2018) to define our Pixel Fidelity Score (PFS), measuring whether encoders attend to contact regions or collapse to background noise (Table 5).

## 3 Task Setup and Data

### 3.1 Task Definition & Dataset

We ground all experiments in the PoseIt dataset (Kanitkar et al., 2022), a multi-sensor recording of a parallel-jaw robot gripper grasping and holding a diverse set of objects across multiple grasp configurations. As seen in figure 2, the dataset comprises 26 everyday objects spanning a wide range of shapes, masses, and surface textures. It includes rigid items such as a flashlight, Rubik's cube,

Figure 2: Dataset Objects. A collection of 26 household objects with a diverse range of size, shape, material, mass, and texture.

and mustard bottle, and deformable or irregular objects such as a bangle, wiper, and tissue paper roll. Each object is grasped under up to 16 distinct pose configurations (varying approach angle and gripper orientation), and at up to four gripper force levels (F5, F15, F40, F80), yielding 1,840 total recorded episodes.

### 3.2 Sensor Modalities

Each episode provides time-aligned recordings from five distinct sources (a sample of which can be seen in figure 3):

- **Tactile images (T):** A GelSight elastomer-style contact sensor captures high-resolution deformation images at approximately 22–26 Hz. Raw frames are baseline-subtracted: each frame has the first frame captured at grasp onset subtracted from it, isolating the contact deformation signal and removing uninformative background texture.
- **RGB images (V):** An external RGB camera records the scene at approximately 10 Hz. Frames are resized to a resolution of $224 \times 224$.
- **Force–Torque (FT):** A six-axis wrench sensor mounted at the end-effector records three forces $(F_x, F_y, F_z)$ and three torques $(T_x, T_y, T_z)$ at approximately 70 Hz. Sensor readings are standardized using per-channel training-split statistics.
- **Gripper state (G):** Left and right finger positions are recorded at approximately 10 Hz, providing a direct measure of the gripper aperture over time.
- **Gripper force command (GF):** A scalar value encoding the commanded gripper force level for the episode (one of F5, F15, F40, F80 Netwons of force), held constant throughout.

### 3.3 Episode Structure and Label Definition

Each episode is divided into five sequential phases: *grasping*, *pose*, *stability*, *retract*, and *release*. Stability outcomes are one of three values per phase: `pass` (object retained), `slip` (lateral displacement), or `drop` (full loss of contact). We frame the prediction target as a **binary classification** task: `pass` maps to label 0 (stable) and `slip` or `drop` map to label 1 (unstable), using the stability phase label as ground truth.

The input window used for prediction is the sum of the *grasping* and *pose* phases, spanning from grasp onset to the start of the stability evaluation. This window has a variable length of roughly 6–20 seconds depending on the episode. Only episodes where the grasp and pose phases both complete without intermediate failure are included, so the model must predict instability from motion cues preceding the shake event rather than reacting to an already-failed grasp.

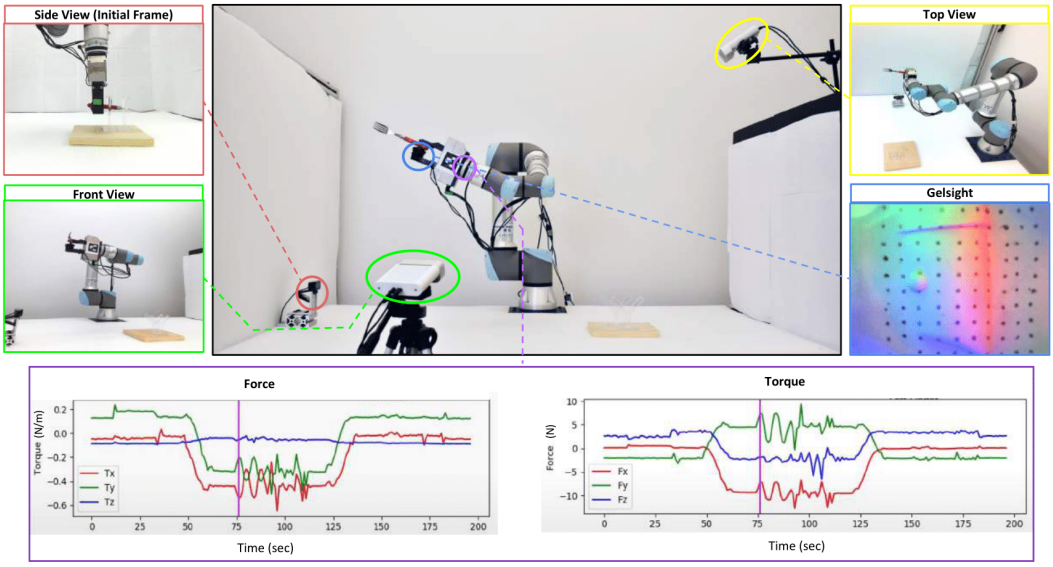



Figure 3: Dataset samples of the multi-modal data, including RGB-D cameras, GelSight tactile sensor, force/torque sensor, and robot trajectory.

### 3.4 Class Imbalance

The dataset exhibits significant class imbalance: approximately 65% of episodes result in a stable grasp (`pass`) and 35% in an unstable outcome (`slip` or `drop`). Because a naive majority-class predictor could achieve over 60% accuracy, accuracy alone is an insufficient evaluation metric. We therefore report accuracy, precision, recall, and F1 scores jointly for all models.

To mitigate the imbalance during training, we optionally employ **Deferred Resampling (DRS)** (Kanitkar et al., 2022): training batches are constructed by thinning the majority (stable) class with a per-batch sub-sampling probability, controlled by a ratio parameter $\sigma$. DRS is activated after an initial warm-up phase so that the model first learns meaningful representations on the full distribution before fine-tuning on a balanced one.

### 3.5 Data Splits

We evaluate under three data splitting protocols to probe different aspects of generalization:

- **Random split (70/15/15):** Episodes are randomly partitioned into train, validation, and test sets, stratified at the episode level. This provides an optimistic estimate of in-distribution performance.
- **Object holdout:** One or more complete object identities are withheld for test, so the model must generalize to unseen object geometries and surface properties.
- **Pose holdout:** One or more grasp pose indices are withheld across all objects, testing generalization to novel grasp configurations.

Unless otherwise stated, all baseline and ablation experiments reported in this work use the random split with a fixed seed for reproducibility.

### 3.6 Sampling and Preprocessing

Sensor streams are sub-sampled to a uniform temporal grid of $F_1 = 1$ image frame per second and $F_2 = 1$ sensor reading per second for both F/T and gripper, producing input tensors of shape $(L, F_1, 3, 224, 224)$ for visual streams and $(L, F_2 \cdot d)$ for tabular streams, where $L$ is the episode length in seconds. Force–torque and gripper features are standardized using training-set mean and variance. All image inputs are resized to $224 \times 224$.

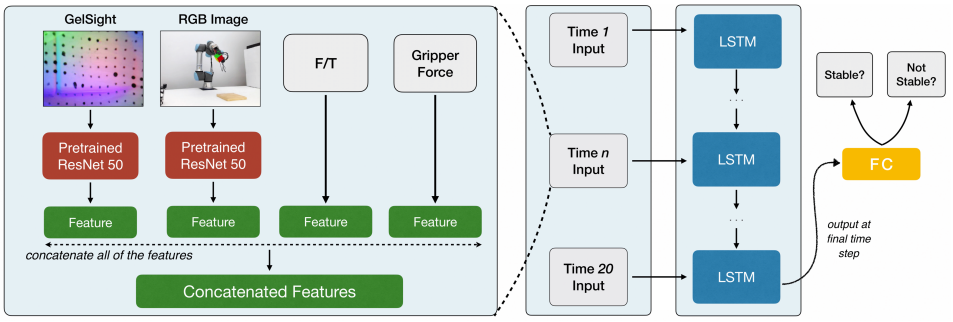



Figure 4: Baseline LSTM model architecture for grasp stability prediction. (excluding all modifications)

## 4 BASELINES

We organize baselines into three tiers of increasing sophistication: unimodal models that assess individual sensing channels in isolation, simple multimodal models that combine modalities via naive concatenation, and competitive baselines that incorporate architectural and training improvements developed over Report 2 and Report 3. All models use the random 70/15/15 split. Table 1 summarizes the results for the multimodal and competitive configurations. The LSTM architecture used to train and validate our baselines can be seen in Figure 4.

### 4.1 UNIMODAL BASELINES

We evaluate each sensing channel independently to characterize its standalone predictive power before any fusion.

**Tactile only (ResNet-50 + MLP)** encodes each GelSight frame with a frozen ImageNet-pretrained ResNet-50 backbone, and classifies with a simple MLP. Despite tactile images directly encoding contact deformation, the model achieves only 34% accuracy and F1 = 0.49, indicating that ImageNet features do not transfer well to GelSight-style contact imagery.

**RGB only (ResNet-50 + MLP)** applies the identical architecture to the external camera stream. It reaches 61% accuracy but F1 = 0.05, collapsing to always predicting *stable*, confirming that slip onset is largely invisible in external RGB.

**Force–Torque only (BiLSTM)** projects the 6-dimensional F/T time-series to 256-d per timestep and processes it with a 2-layer BiLSTM. Despite the low input dimensionality it achieves an accuracy of 48% and F1 = 0.52, competitive with the tactile baseline, highlighting the signal value of wrench readings.

### 4.2 SIMPLE MULTIMODAL BASELINES

**V+T (ResNet-50, BiLSTM)** concatenates ResNet-50 features from both visual streams per timestep and processes them with a 2-layer BiLSTM. Combining the two visual modalities with temporal context yields a substantial gain: 77.6% accuracy and F1 = 0.710, far exceeding either visual stream alone.

**V+T+FT+G+GF naive concatenation** extends the above to all five modalities by simply appending raw features before the BiLSTM. Accuracy collapses to 53.4% (F1 = 0.561). The degradation is caused by severe dimensional dominance: the visual streams contribute 2048-d each while F/T, gripper, and gripper force contribute 12, 6, and 1 dimensions respectively, causing the model to effectively ignore all non-visual signals.

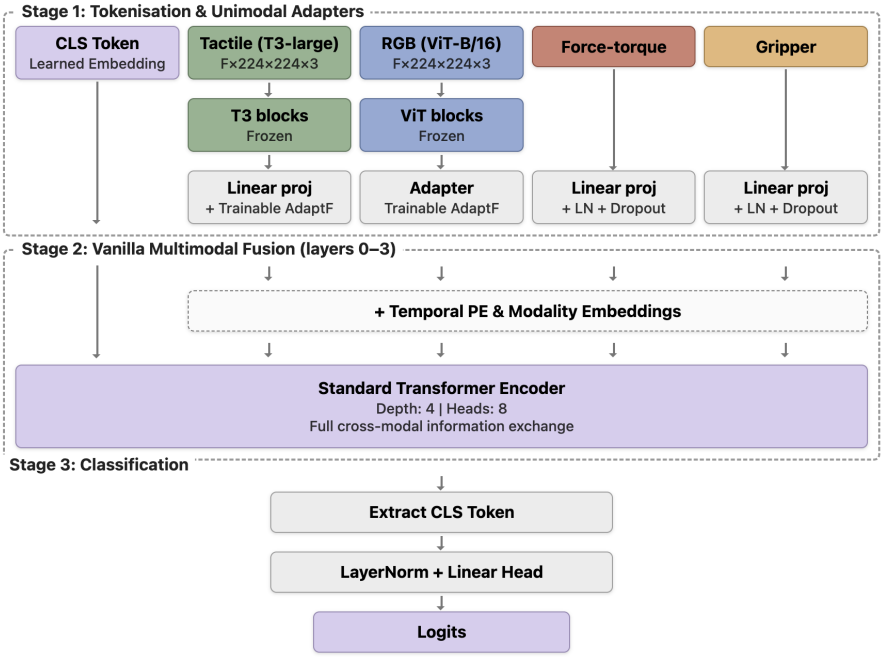



Figure 5: Vanilla Transformer Architecture

### 4.3 Competitive Baselines

**V+T+FT+G+GF with balanced projections (no DRS)** applies per-modality linear projections before concatenation ($p_V = p_T = 512$, $p_{FT} = 128$, $p_G = 64$, $p_{GF} = 32$) to improve feature dimension imbalance. This recovers performance to 80.9% accuracy and F1 = 0.752, confirming dimensional dominance as the primary failure mode of naive five-modal fusion.

**V+T+FT+G+GF with CLIP ViT-L/14 + T3-large (no DRS)** replaces the ResNet-50 backbone with OpenCLIP ViT-L/14 (Radford et al., 2021) for RGB and T3-large (Zhao et al., 2024) for tactile; with the motivation of using encoders with richer, task-appropriate priors. Combined with balanced projections, this achieves 83.5% accuracy and F1 = 0.800, the strongest BiLSTM-family result.

**Vanilla Multimodal Transformer** is one of our proposed models from Report 2 (see figure 5). After running experiments, we treat it as another competitive baseline for the proposed Multimodal Bottleneck Transformer architecture. It treats all modalities jointly in a single shared attention space. Modality and temporal positional embeddings distinguish tokens before they are processed by $L$ stacked transformer encoder blocks. The V+T+FT configuration reaches 79.4% accuracy and F1 = 0.767; extending to all five modalities (V+T+FT+G+GF) yields 76.7% accuracy and F1 = 0.777. The slight drop when adding G and GF echoes the modality dominance pattern: without a controlled fusion mechanism, low-dimensional streams can be crowded out even inside self-attention.

### 4.4 Evaluation Metrics

We evaluate all models using four complementary metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \qquad \text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \qquad F_1\text{Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

Given the class imbalance in the dataset, we treat $F_1$ as the primary ranking metric: it captures the trade-off between false alarms and missed slips. Accuracy is reported alongside $F_1$ for completeness.

Table 1: Baseline and their corresponding test metrics. Proposed MBT results are reported and discussed in section 6.

| **Configuration** | **Accuracy (%)** ↑ | **F1** ↑ |
|---|---|---|
| *Unimodal* | | |
| T only (ResNet-50 + MLP) | 34.0 | 0.492 |
| V only (ResNet-50 + MLP) | 60.6 | 0.051 |
| FT only (BiLSTM) | 48.9 | 0.524 |
| *Simple Multimodal* | | |
| V+T (ResNet-50, BiLSTM) | 77.6 | 0.710 |
| V+T+FT+G+GF naive concat (BiLSTM) | 53.4 | 0.561 |
| *Competitive* | | |
| V+T+FT+G+GF + balanced proj., no DRS (BiLSTM) | 80.9 | 0.752 |
| V+T+FT+G+GF + CLIP/T3, no DRS (BiLSTM) | 83.5 | 0.800 |
| Vanilla Transformer (V+T+FT) | 79.4 | 0.767 |
| Vanilla Transformer (V+T+FT+G+GF) | 76.7 | 0.777 |
| ***Proposed (Multimodal Bottleneck Transformer)*** | | |

## 5 PROPOSED MODEL

The model that we propose is based off of the Multimodal Bottleneck Transformer (MBT) Nagrani et al. (2022), originally designed for two-modality audio-visual fusion. MBT restricts cross-modal information exchange to a small set of learned bottleneck latent tokens, forcing the model to compress and share only the most task-relevant information across modalities. We extend this framework from two modalities to five for grasp stability prediction: tactile images (T), RGB images (V), force-torque (FT), gripper state (G), and scalar gripper force (GF). The model overview is shown in figure 6.

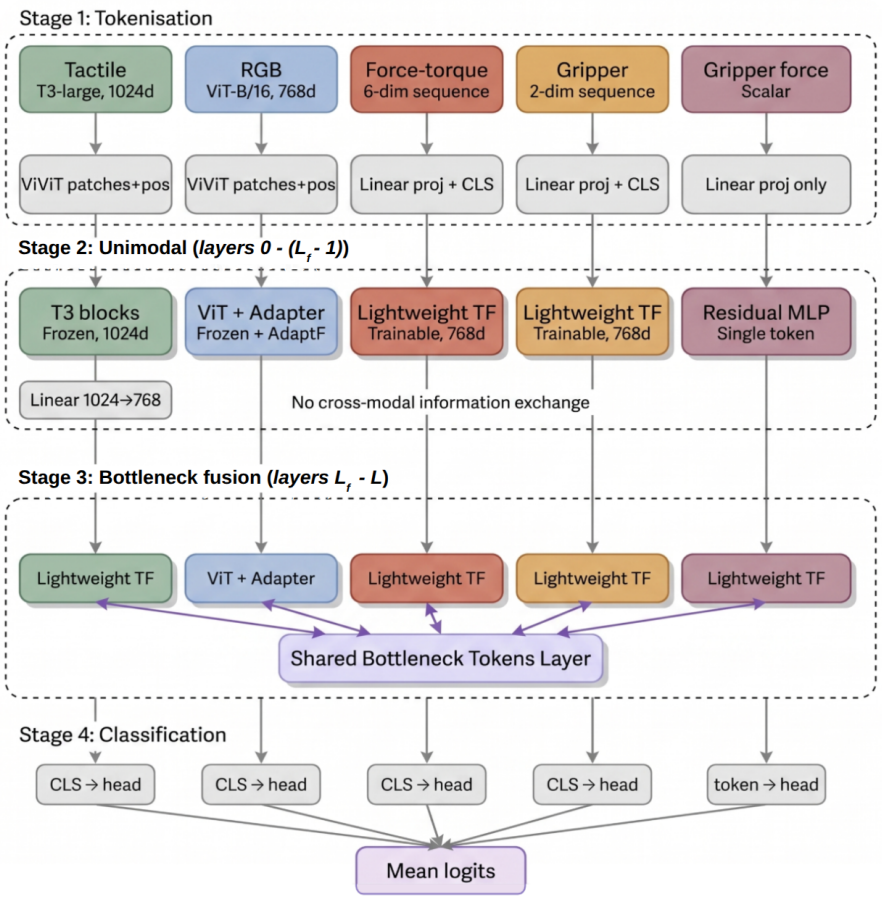



Figure 6: Multimodal Bottleneck Transformer Architecture

**Tokenization.** The two visual streams, tactile and RGB, are tokenized into patch sequences using their respective pretrained backbones (T3-large at 1024-d for tactile; ViT-Base/16 at 768-d for RGB), with ViViT-style factored spatial and temporal positional embeddings and a prepended CLS token. Both are uniformly subsampled to a maximum of $F = 8$ frames. Force-torque and gripper sequences are projected to 768-d via a linear layer and layer normalization, with a learned CLS token and positional embeddings. Gripper force, a single scalar, is projected to one 768-d token.

**Unimodal processing (layers $0$ to $L_f - 1$).** Following MBT's mid-fusion strategy, the first $L_f$ layers process each modality independently via standard transformer layers. The tactile and RGB streams use their respective frozen pretrained blocks, with AdaptFormer bottleneck adapters Chen et al. (2022) added in parallel to each RGB block's FFN for parameter-efficient fine-tuning. Force-torque and gripper use lightweight trainable transformer blocks (4 heads, $2\times$ MLP ratio). Gripper force uses a residual MLP. At the fusion boundary, the tactile stream is linearly projected from 1024-d to 768-d so all streams share a common dimensionality.

**Bottleneck fusion (layers $L_f$ to $L-1$)).** A set of $B$ shared bottleneck tokens, initialised from $\mathcal{N}(0, 0.02)$, serve as the sole conduit for cross-modal information. At each fusion layer, the bottleneck tokens are concatenated with each stream's sequence and processed via self-attention:

$$[\mathbf{z}_i^{l+1} \,\|\, \hat{\mathbf{z}}_{\text{fsn},i}^{l+1}] = \text{Transformer}([\mathbf{z}_i^l \,\|\, \mathbf{z}_{\text{fsn}}^l]; \theta_i), \tag{3}$$

$$\mathbf{z}_{\text{fsn}}^{l+1} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \hat{\mathbf{z}}_{\text{fsn},i}^{l+1}, \tag{4}$$

where $i$ indexes the active modalities $\mathcal{M} \subseteq \{\text{T, V, FT, G, GF}\}$ and $\|$ denotes concatenation. Each modality produces a temporary bottleneck update $\hat{\mathbf{z}}_{\text{fsn},i}$, and the shared bottleneck state is obtained by averaging across modalities. RGB fusion blocks reuse the pretrained ViT layers with adapters; all other streams use lightweight trainable blocks.

**Classification.** Each modality's CLS token is passed through a modality-specific layer norm and linear head, and the final prediction is the mean of all per-modality logits:

$$\hat{y} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbf{W}_i \, \text{LN}_i(\mathbf{z}_{\text{cls-}i}^L). \tag{5}$$

### 5.1 Loss functions

We train with binary cross-entropy loss, as grasp stability is framed as a binary classification task (stable vs. unstable). Given the raw logit $\hat{y}$ and ground-truth label $y \in \{0, 1\}$:

$$\mathcal{L} = -\big[y \log \sigma(\hat{y}) + (1 - y) \log(1 - \sigma(\hat{y}))\big], \tag{6}$$

where $\sigma$ denotes the sigmoid function.

### 5.2 Changes to training data

We disable Deferred Resampling (DRS) for all MBT experiments because activating DRS mid-training introduces a sudden distribution shift that destabilizes optimization, producing sharp loss spikes and accuracy drops shortly after the resampling flag is triggered (Table 6). Enabling DRS from epoch zero avoids this discontinuity but substantially reduces the already limited effective dataset size. We instead rely on balanced per-modality projections and OGM to manage imbalance, which yields stronger results in practice. No other changes to the training data were made.

### 5.3 Hyperparameters and their effects

We ablate two key architectural hyperparameters of the MBT: the number of bottleneck tokens $B \in \{4, 8, 16\}$ and the fusion layer depth $L_f \in \{2, 4, 8\}$, across four modality subsets. All runs in this section disable OGM; results are summarized in Table 2.

**Effect of bottleneck size $B$.** Increasing $B$ does not yield consistent gains. For T,V,FT,G,GF, the best result (82.6% accuracy, F1 = 0.767) is achieved at $B = 8, L_f = 8$, while $B = 16$ at the same fusion depth slightly underperforms. For T,V,FT, the highest accuracy (84.3%) occurs at $B = 16, L_f = 2$, yet F1 peaks at $B = 8, L_f = 4$ (0.767). This inconsistency suggests that a larger bottleneck capacity does not straightforwardly improve cross-modal information sharing, likely because the bottleneck tokens must simultaneously compress representations from streams of very different dimensionality and sequence length. A bottleneck that is too large may allow modality-specific information to pass through without being meaningfully distilled into a shared representation.

**Effect of fusion depth $L_f$.** Later fusion (larger $L_f$) gives each modality more layers of unimodal processing before cross-modal exchange begins. For visual-dominant sets such as V,FT,GF, later fusion tends to be beneficial — the model at $L_f = 8$ achieves competitive results across all $B$ values — consistent with the intuition that pretrained visual encoders benefit from deeper unimodal refinement before contributing to the shared bottleneck. The pattern is less clean for modality sets that include tactile, where the optimal $L_f$ varies with $B$, suggesting that tactile representations may benefit from earlier cross-modal grounding rather than extended unimodal processing in isolation.

Table 2: MBT hyperparameter ablations (OGM disabled). Absent rows indicate failed runs.

| Modalities | $B$ | $L_f$ | Test Acc (%) ↑ | Test F1 ↑ |
|---|---|---|---|---|
| T,V,FT,G,GF | 4 | 2 | 75.3 | 0.714 |
| T,V,FT,G,GF | 4 | 4 | 78.1 | 0.711 |
| T,V,FT,G,GF | 4 | 8 | 75.8 | 0.726 |
| T,V,FT,G,GF | 8 | 2 | 79.8 | 0.763 |
| T,V,FT,G,GF | 8 | 4 | 75.3 | 0.699 |
| T,V,FT,G,GF | 8 | 8 | 82.6 | 0.767 |
| T,V,FT,G,GF | 16 | 2 | 80.3 | 0.748 |
| T,V,FT,G,GF | 16 | 4 | 78.1 | 0.769 |
| T,V,FT,G,GF | 16 | 8 | 79.2 | 0.748 |
| T,V,FT | 4 | 2 | 77.0 | 0.687 |
| T,V,FT | 4 | 4 | 79.8 | 0.743 |
| T,V,FT | 4 | 8 | 70.2 | 0.686 |
| T,V,FT | 8 | 2 | 78.7 | 0.765 |
| T,V,FT | 8 | 4 | 78.7 | 0.721 |
| T,V,FT | 8 | 8 | 71.9 | 0.658 |
| T,V,FT | 16 | 2 | 84.3 | 0.741 |
| T,V,FT | 16 | 4 | – | – |
| T,V,FT | 16 | 8 | 77.5 | 0.710 |
| V,FT,GF | 4 | 2 | 83.1 | 0.773 |
| V,FT,GF | 4 | 4 | 75.8 | 0.711 |
| V,FT,GF | 4 | 8 | 80.9 | 0.734 |
| V,FT,GF | 8 | 2 | 82.6 | 0.744 |
| V,FT,GF | 8 | 4 | – | – |
| V,FT,GF | 8 | 8 | 80.9 | 0.767 |
| V,FT,GF | 16 | 2 | 74.7 | 0.667 |
| V,FT,GF | 16 | 4 | – | – |
| V,FT,GF | 16 | 8 | 82.6 | 0.786 |
| T,FT,GF | 4 | 2 | 65.2 | 0.644 |
| T,FT,GF | 4 | 4 | – | – |
| T,FT,GF | 4 | 8 | 69.7 | 0.682 |
| T,FT,GF | 8 | 2 | 70.2 | 0.649 |
| T,FT,GF | 8 | 4 | – | – |
| T,FT,GF | 8 | 8 | 65.2 | 0.622 |
| T,FT,GF | 16 | 2 | 70.8 | 0.708 |
| T,FT,GF | 16 | 4 | 71.9 | 0.653 |
| T,FT,GF | 16 | 8 | 65.7 | 0.596 |

**Modality subset effects.** The T,FT,GF results (65.2–71.9% accuracy) are the weakest across configurations, which we attribute primarily to the absence of RGB rather than any deficiency of tactile itself. As the unimodal baselines show, the visual stream carries strong scene-level priors that the other modalities cannot fully substitute. Notably, V,FT,GF remains competitive (74.7–83.1%), but the full five-modality set T,V,FT,G,GF consistently achieves the most stable results, suggesting that tactile contributes meaningfully to fusion even when its standalone signal is modest. The proprioceptive streams (G, GF), while low-dimensional, appear to further stabilize training when combined with richer visual and tactile representations.

## 6 RESULTS

Table 3 reports test-set performance across all evaluated architectures. We use F1 as the primary metric given the class imbalance in PoseIt (∼70% stable, ∼30% unstable), where raw accuracy is a poor proxy for model quality.

Table 3: Test-set results for the proposed architecture

| Configuration | Acc (%) ↑ | F1 ↑ |
|---|---|---|
| *Multimodal Bottleneck Transformer (MBT)* | | |
| V, T, FT, GF, G | 82.6% | 0.767 |
| V, T FT | **84.3%** | 0.741 |
| V, FT, GF | 83.7% | **0.782** |
| T, FT, GF | 73.0% | 0.676 |
| *Vanilla Transformer* | | |
| V, T, FT, GF, G | 76.7% | 0.777 |
| V, T, FT | 79.4% | 0.767 |
| V, FT, GF | 79.7% | 0.730 |
| T, FT, GF | 70.6% | 0.683 |

**Proposed MBT Architecture.** The MBT achieves strong results across all modality subsets, with the best configurations outperforming all Vanilla Transformer baselines on both metrics. The full five-modality set T,V,FT,G,GF reaches 82.6% accuracy and F1 = 0.767 ($B$=8, $L_f$=8, no OGM), while V,FT,GF achieves the highest F1 of 0.782 at 83.7% accuracy ($B$=4, $L_f$=2, with OGM). Notably, T,V,FT yields the highest raw accuracy of 84.3% ($B$=16, $L_f$=2, no OGM), though at a slightly lower F1 of 0.741, suggesting the model favors the majority stable class in this configuration. The weakest subset remains T,FT,GF (73.0%, F1 = 0.676), which we attribute to the absence of RGB rather than any deficiency of the tactile stream. This is also a conclusion supported by the fact that V,FT,GF, which drops tactile instead, remains highly competitive.

**Comparison to Vanilla Transformer.** The MBT outperforms the Vanilla Transformer on accuracy across all four modality subsets, with gains of up to 5.9 percentage points (V,FT,GF: 83.7% vs. 79.7%). However, the picture is more nuanced on F1: for T,V,FT,G,GF and T,V,FT, the Vanilla Transformer achieves slightly higher F1 (0.777 vs. 0.767 and 0.767 vs. 0.741 respectively), suggesting that MBT's bottleneck mechanism, while improving overall accuracy, can occasionally bias predictions toward the majority stable class in modality-rich configurations. For V,FT,GF and T,FT,GF, MBT improves on both metrics, indicating that the bottleneck fusion is most beneficial when modalities are heterogeneous and cross-modal interference is otherwise high.

**Comparison to Recurrent Baselines.** The strongest BiLSTM baseline (CLIP ViT-L/14 + T3-large, 83.5% accuracy, F1 = 0.800) remains competitive with and slightly ahead of the best MBT configuration on F1. We attribute this to the limited dataset size of $\sim$1700 episodes: BiLSTMs benefit from a stronger inductive bias on short temporal sequences and require far fewer parameters to fit, whereas transformer-based models typically require substantially more data to realize their representational capacity. Despite this, MBT operates on a more principled fusion mechanism and demonstrates stronger modality balance as evidenced by the discrepancy ratio analysis in Table 4.

# 7 ANALYSIS

## 7.1 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:

$$s_i^u = \begin{cases} \sigma(\hat{y}_i^u) & y_i = 1 \\ 1 - \sigma(\hat{y}_i^u) & y_i = 0 \end{cases} \tag{7}$$

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

$$\rho^u = \frac{\sum_{i \in \mathcal{B}} s_i^u}{\frac{1}{|\mathcal{M}|-1} \sum_{v \neq u} \sum_{i \in \mathcal{B}} s_i^v} \tag{8}$$

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

Table 4 shows consistent modality imbalance across configurations. In nearly all settings, the visual stream dominates ($\rho_V > 1$) while the force-torque stream is under-optimized ($\rho_{FT} < 1$), indicating the model disproportionately relies on visual features at the expense of proprioceptive ones. OGM partially corrects this imbalance and yields accuracy and F1 improvements most reliably for the V,FT,GF and T,V,FT,G,GF modality sets. However, gains are less consistent for T,V,FT and T,FT,GF, where OGM occasionally reduces performance relative to the no-OGM baseline, suggesting that the tactile stream interacts with gradient rescaling in less predictable ways. Notably, neither increasing bottleneck size $B$ nor fusion depth $L_f$ monotonically improves performance in either case, suggesting that architectural capacity alone is insufficient without proper gradient balancing.

Figure 7 visualises these trends across all four modality sets, averaged over all $B$ and $L_f$ configurations. The top row confirms that OGM produces the most reliable accuracy and F1 gains for V,FT,GF and T,V,FT,G,GF, while improvements for the remaining sets are marginal or absent. The bottom panel reveals the underlying imbalance structure: $\rho_V$ sits well above unity in every group containing vision, peaking near 2.6 for V,FT,GF, indicating that visual dominance persists even after OGM is applied. Where vision is absent, as in T,FT,GF, the tactile stream fills the dominant role ($\rho_T > 1$), pointing to a broader tendency for the model to anchor on its strongest available signal rather than leverage complementary modalities. Throughout, $\rho_{FT}$ remains the most consistently sub-unity ratio across all groups, suggesting that force-torque features are chronically undertrained — a pattern that may reflect limitations in the force-torque encoder itself rather than a gradient balancing issue that OGM alone can resolve.

**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 $\pi$ be the pixel ordering induced by GradCAM importance (descending). Starting from a Gaussian-blurred (information-destroyed) version of the image, we progressively reveal pixels in order $\pi$, 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

$$\text{PFS} = \text{AUC}\big(\{c_{\text{ins}}(k)\}_{k=0}^{100}\big), \tag{9}$$

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.

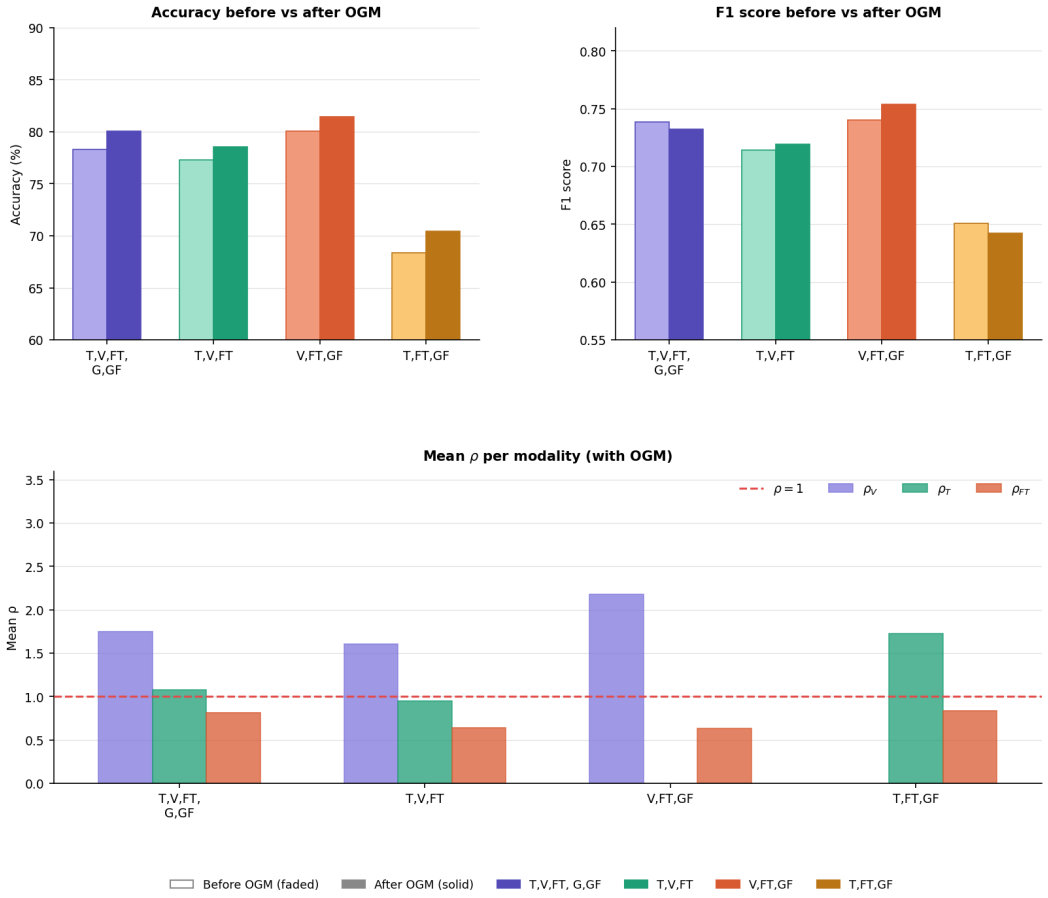



Figure 7: Mean accuracy, F1 score, and per-modality discrepancy ratios ($\rho$) across four modality sets, averaged over all batch size ($B$) and fusion layer ($L_f$) configurations. Top row: accuracy and F1 before (faded) vs. after (solid) OGM. Bottom: mean $\rho$ for visual (V), text (T), and fine-tuned (FT) modalities with OGM applied; dashed red line marks $\rho = 1$.

We apply EigenGradCAM Muhammad & Yeasin (2020) to the last unimodal transformer block of each visual stream (pre-fusion), which produces sharper spatial maps than post-fusion layers.

Table 5 compares PFS across LSTM-family baselines from R3 and our final MBT model. The LSTM baselines collapse to near-chance PFS ($\approx 0.50$) as additional modalities are added, indicating that the RGB encoder abandons task-relevant contact regions under joint training. The two-modal Tactile+RGB LSTM retains the highest LSTM PFS (0.749), but our MBT model substantially improves on this: **0.878** for tactile and **0.856** for RGB, suggesting that MBT's independent per-stream processing in the unimodal layers prevents visual attention from collapsing to background regions. This failure mode was consistently observed in LSTM-based fusion where from the first timestep dominant modalities corrupt the visual gradient signal.

Figure 8 shows EigenGradCAM activations for the T3-large tactile and CLIP ViT-L/14 RGB encoders across seven validation samples. The T3-large backbone consistently localizes attention to physically meaningful contact regions. The tactile maps consistently highlight the fingertip contact patch and elastomer deformation zone, while the RGB maps attend to the gripper-object interface rather than the background. This is in contrast to the LSTM baseline, whose GradCAM maps scatter across the image background once F/T is added.

Table 4: Per-modality discrepancy ratios $\rho^u$ before and after OGM. $\rho^u > 1$ indicates dominance; $\rho^u < 1$ indicates under-optimization.

| | | | Without OGM | | With OGM | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Modalities | $B$ | $L_f$ | Acc (%) ↑ | F1 ↑ | $\rho_V$ | $\rho_T$ | $\rho_{FT}$ | Acc (%) ↑ | F1 ↑ |
| T,V,FT,G,GF | 4 | 2 | 75.3 | 0.714 | 1.677 | 0.727 | 0.743 | 80.9 | 0.742 |
| T,V,FT,G,GF | 4 | 4 | 78.1 | 0.711 | 1.738 | 1.601 | 0.904 | 80.3 | 0.741 |
| T,V,FT,G,GF | 4 | 8 | 75.8 | 0.726 | 1.637 | 1.488 | 0.678 | 82.6 | 0.763 |
| T,V,FT,G,GF | 8 | 2 | 79.8 | 0.763 | 1.979 | 0.735 | 1.013 | 80.3 | 0.724 |
| T,V,FT,G,GF | 8 | 4 | 75.3 | 0.699 | 2.127 | 1.068 | 0.825 | 80.3 | 0.711 |
| T,V,FT,G,GF | 8 | 8 | 82.6 | 0.767 | 1.455 | 0.765 | 0.847 | – | – |
| T,V,FT,G,GF | 16 | 2 | 80.3 | 0.748 | 1.941 | 0.923 | 0.762 | 77.5 | 0.730 |
| T,V,FT,G,GF | 16 | 4 | 78.1 | 0.769 | 1.873 | 1.387 | 0.708 | 78.7 | 0.716 |
| T,V,FT,G,GF | 16 | 8 | 79.2 | 0.748 | 1.338 | 1.033 | 0.842 | – | – |
| T,V,FT | 4 | 2 | 77.0 | 0.687 | 1.614 | 0.840 | 0.694 | 75.8 | 0.691 |
| T,V,FT | 4 | 4 | 79.8 | 0.743 | 1.858 | 0.674 | 0.726 | 77.0 | 0.637 |
| T,V,FT | 4 | 8 | 70.2 | 0.686 | 1.672 | 0.934 | 0.585 | 79.8 | 0.743 |
| T,V,FT | 8 | 2 | 78.7 | 0.765 | 1.726 | 0.718 | 0.750 | 79.8 | 0.746 |
| T,V,FT | 8 | 4 | 78.7 | 0.721 | 1.028 | 1.538 | 0.584 | 82.6 | 0.767 |
| T,V,FT | 8 | 8 | 71.9 | 0.658 | 1.624 | 0.981 | 0.573 | 77.5 | 0.730 |
| T,V,FT | 16 | 2 | 84.3 | 0.741 | 1.268 | 1.021 | 0.755 | 78.7 | 0.716 |
| T,V,FT | 16 | 4 | – | – | 1.459 | 1.089 | 0.582 | 78.7 | 0.716 |
| T,V,FT | 16 | 8 | 77.5 | 0.710 | 2.230 | 0.750 | 0.500 | 77.0 | 0.728 |
| V,FT,GF | 4 | 2 | 83.1 | 0.773 | 2.872 | – | 0.763 | 83.7 | 0.782 |
| V,FT,GF | 4 | 4 | 75.8 | 0.711 | 1.965 | – | 0.654 | 78.1 | 0.723 |
| V,FT,GF | 4 | 8 | 80.9 | 0.734 | 1.355 | – | 0.780 | 80.9 | 0.754 |
| V,FT,GF | 8 | 2 | 82.6 | 0.744 | 1.514 | – | 0.741 | 82.0 | 0.750 |
| V,FT,GF | 8 | 4 | – | – | 2.058 | – | 0.777 | 80.9 | 0.750 |
| V,FT,GF | 8 | 8 | 80.9 | 0.767 | 3.319 | – | 0.263 | 83.1 | 0.786 |
| V,FT,GF | 16 | 2 | 74.7 | 0.667 | 2.418 | – | 0.664 | 81.5 | 0.759 |
| V,FT,GF | 16 | 4 | – | – | 1.494 | – | 0.584 | 82.0 | 0.750 |
| V,FT,GF | 16 | 8 | 82.6 | 0.786 | 2.624 | – | 0.495 | 80.9 | 0.730 |
| T,FT,GF | 4 | 2 | 65.2 | 0.644 | – | 1.595 | 0.809 | 67.4 | 0.623 |
| T,FT,GF | 4 | 4 | – | – | – | 1.666 | 0.680 | 69.7 | 0.649 |
| T,FT,GF | 4 | 8 | 69.7 | 0.682 | – | 2.571 | 0.671 | 71.9 | 0.662 |
| T,FT,GF | 8 | 2 | 70.2 | 0.649 | – | 1.923 | 0.713 | 68.5 | 0.627 |
| T,FT,GF | 8 | 4 | – | – | – | 2.543 | 0.616 | 68.5 | 0.627 |
| T,FT,GF | 8 | 8 | 65.2 | 0.622 | – | 1.100 | 1.527 | 72.5 | 0.657 |
| T,FT,GF | 16 | 2 | 70.8 | 0.708 | – | 1.491 | 0.773 | 71.3 | 0.643 |
| T,FT,GF | 16 | 4 | 71.9 | 0.653 | – | 1.838 | 0.608 | 71.3 | 0.617 |
| T,FT,GF | 16 | 8 | 65.7 | 0.596 | – | 0.810 | 1.109 | 73.0 | 0.676 |

Table 5: Pixel Fidelity Score (PFS, Insertion AUC) across model configurations. MBT scores use EigenGradCAM on unimodal blocks; LSTM scores use GradCAM on ResNet-50. Higher is better; 0.5 = random.

| **Model** | **Acc (%) ↑** | **PFS (Tac) ↑** | **PFS (RGB) ↑** |
|---|---|---|---|
| Tactile + RGB (LSTM) | 74.8 | 0.749 | 0.749 |
| Tactile + F/T (LSTM) | 54.2 | 0.502 | — |
| Tactile + RGB + F/T (LSTM) | 51.1 | 0.499 | 0.499 |
| Tactile + RGB + F/T + GF (LSTM) | 59.2 | 0.501 | 0.501 |
| Tactile + RGB + F/T + GF + G (LSTM) | 48.5 | 0.501 | 0.501 |
| **MBT (V+T+FT+G+GF)** | **83.2** | **0.878** | **0.856** |

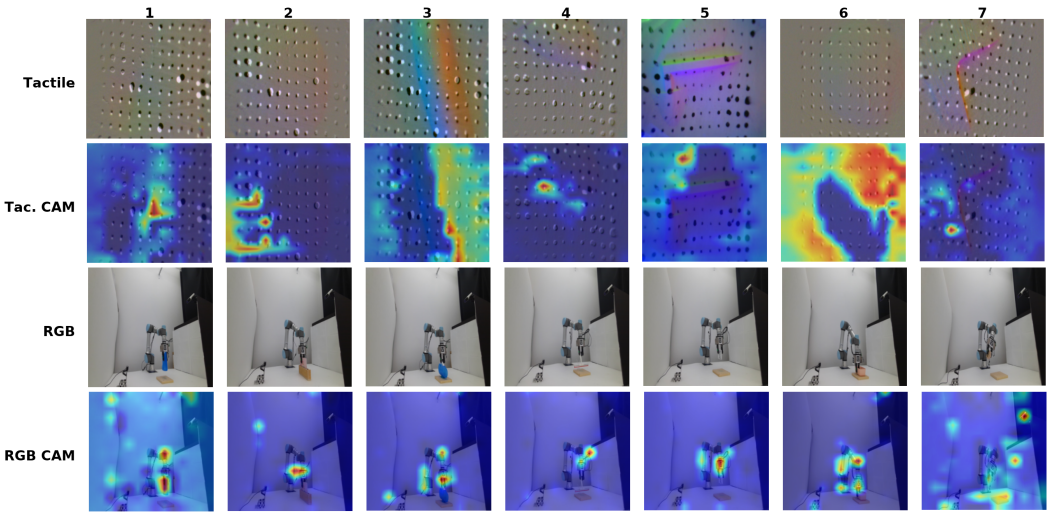
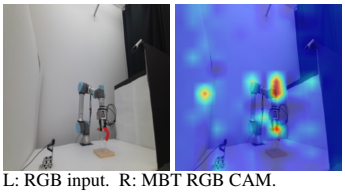
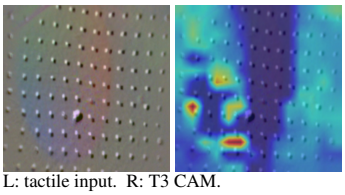
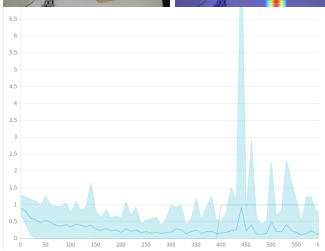



Figure 8: EigenGradCAM visualizations for seven validation samples (columns 1–7). **Row 1**: raw tactile frame (GelSight). **Row 2**: tactile saliency map (T3-large backbone). **Row 3**: raw RGB frame. **Row 4**: RGB saliency map (CLIP ViT-L/14 backbone). Warm colors (red/yellow) indicate high activation; cool colors (blue) indicate low activation.

### 7.2 Qualitative Analysis and Examples

| Observation | Evidence | Analysis |
|---|---|---|
| Balanced projection recovers dominance | Naive: 53.4%, F1=0.561<br>Proj.: 80.9%, F1=0.752 | Projecting each modality to a shared dim restores and surpasses the two-modal baseline. |
| Visual attention collapse in LSTM; Recovery in MBT |  <br>L: RGB input. R: MBT RGB CAM.<br>PFS T+V LSTM: 0.749<br>PFS T+V+FT LSTM: 0.499<br>PFS MBT 5-modal: **0.856** | Adding F/T collapses LSTM visual PFS; Grad-CAM shifts to background walls. For most samples, MBT's preserves visual attention on the gripper and object in the RGB Frame. |
| T3 + ViViT v.s. ResNet activation quality |  <br>L: tactile input. R: T3 CAM.<br>ResNet PFS: ≈0.499<br>T3 PFS: **0.878**, CLIP PFS: **0.856** | ResNet activates on background; T3 focuses on contact boundary and dot-array distortions diagnostic of slip. Encoder choice drives attention quality more than architecture. |
| RGB encoder conflates object and background under low color contrast |  | The grasped object is a white paper towel roll held in front of a white wall. The model fails to localize it and instead activates on the wooden block on the floor. This was observed on multiple cases when object and background share similar color and texture. |
| DRS makes training unstable |  <br>The spike occurs soon after `drs_active` flag becomes active. | Activating DRS mid-training introduces a sudden distribution shift that destabilizes optimization. Enabling it from epoch zero avoids this discontinuity but substantially reduces the effective dataset size, which is already limited at ∼1,700 episodes. We therefore disable DRS entirely in favor of balanced projections and OGM. |
| Tactile sensing is not a very important modality in determining whether the object grasp is stable | Accuracy of Vanilla Transformer with V, FT, GF = 79.7%, with T, FT, GF = 70.6%. MBT ablations show this too: V,FT,GF (74.7–83.1%) matches the full set, while T,FT,GF collapses to 65.2–71.9%. | We hypothesize that this is the case because the Tactile Encoder used was not extracting useful features — the image modality was much more useful as the information to determine whether a grasp is stable was readily available in the extracted features. |
| OGM partially corrects modality imbalance but cannot fully resolve it | OGM improves accuracy most reliably for V,FT,GF and T,V,FT,G,GF, but hurts or shows no gain for T,V,FT and T,FT,GF. $\rho_V$ remains $> 1$ even after OGM is applied (peaking near 2.6 for V,FT,GF). | OGM operates on gradient magnitudes during training, but representational damage from early visual dominance is already encoded before reweighting can intervene. This suggests the imbalance is partly structural. Visual streams contribute far longer token sequences through the bottleneck than proprioceptive ones, and cannot be fully corrected by gradient rescaling alone. |

| **Observation** | **Evidence** | **Analysis** |
|---|---|---|
| The vanilla transformer was unable to reach the accuracy of the recurrent methods | Vanilla Transformer accuracy: 79.7%<br>MBT accuracy: 82.6%<br>Recurrent accuracy: 83.5% | 1700 samples was not enough to train data-hungry transformer-based models model fully with enough generalization for other objects and poses. Moreover, training for too long on the current dataset begins to overfit the model to the training set. |
| Optimal fusion depth $L_f$ is inconsistent across modality subsets and bottleneck sizes | For T,V,FT: best F1 at $L_f = 4$ ($B = 8$, F1$= 0.767$) but worst accuracy at $L_f = 8$ ($B = 4, 70.2\%$). For V,FT,GF: $L_f = 8$ competitive across all $B$ values. | Pretrained visual streams benefit from deeper unimodal refinement before fusion, while tactile representations appear to benefit from earlier cross-modal grounding. The inconsistency suggests $L_f$ should be tuned per modality subset rather than treated as a global hyperparameter. |
| G and GF have mixed effects on performance despite low dimensionality | Vanilla Transformer increases from F1 $= 0.767$ (V,T,FT) to F1 $= 0.777$ (V,T,FT,G,GF) but accuracy drops from 79.4% to 76.7%. MBT full five-modal set achieves most stable results across ablations relative to modality subsets without G and GF. | G and GF appear to improve minority class recall (reflected in higher F1) at a slight cost to overall accuracy, suggesting they provide episode-level context that helps distinguish unstable grasps. However, their inconsistent effect across architectures makes it difficult to draw strong conclusions about their contribution. |
| MBT achieves higher accuracy than Vanilla Transformer but lower F1 in modality-rich configurations | T,V,FT,G,GF: MBT 82.6% acc / F1$=$ $0.767$ vs. Vanilla 76.7% acc / F1$=$ $0.777$. T,V,FT: MBT 84.3% acc / F1$= 0.741$ vs. Vanilla 79.4% acc / F1$= 0.767$. | The cause is not fully diagnosed. One hypothesis is that the bottleneck compresses cross-modal information in a way that favors the majority stable class signal when more modalities are present. Another is that MBT's larger parameter count makes it more prone to majority class overfitting on the limited $\sim$1700 episode dataset without DRS. |
| Gelsight sensors change degrade over time, changing the properties of the sensor over repeated runs of hardware collection | Gelsight sensors are polymer based. As more samples are collected by the hardware, the top layer of the polymer gets degraded, changing the properties of the sensor as time goes on. | Elastomer wear shifts the sensor's deformation statistics over time, causing the model to potentially overfit to a specific wear regime rather than generalizable contact geometry. Periodic recalibration or domain adaptation between wear states could mitigate this. |

Table 6: Qualitative failure analysis.

## 8 FUTURE WORK AND LIMITATIONS

- **Dataset scale and diversity:** The PoseIt dataset, while carefully constructed, contains only 493 episodes across 26 object types, creating a tight data budget for fine-tuning large pre-trained transformers such as ViT-Base and T3-large. This is compounded by a significant class imbalance: approximately 70% of episodes are stable grasps, leaving only $\sim$30% as the hard unstable cases that the model most needs to learn. While deferred resampling partially mitigates this, models trained at this scale risk over-fitting to object-specific appearance and grasping dynamics rather than learning generalizable stability cues.
- **Single sensor and robot configuration:** All data originates from a single GelSight sensor and one parallel-jaw gripper configuration on one robot platform. GelSight produces elastomer-deformation contact images that are visually and dimensionally distinct from sensors such as Digit360 or TACTO-based sensors. Even with the T3-large tactile foundation encoder, the learned representations are biased toward the geometric and photometric properties of GelSight, and re-using this model on a different tactile sensor would require substantial retraining or domain adaptation.
- **Unequal compression through the bottleneck:** RGB and tactile streams each contribute long patch-token sequences to every fusion layer, while force–torque, gripper, and gripper-force streams contribute far fewer tokens. When all streams compete through the same small set of bottleneck tokens, visual streams exert greater influence over the shared representation simply by volume. This asymmetry is structurally analogous to the dimensional dominance observed in naive concatenation baselines, suggesting that per-stream bottleneck budgets or stream-aware attention masking may be necessary for truly balanced fusion.
- **Episode-level prediction only:** The model produces a single binary label per grasp episode rather than a per-timestep stability estimate. This design cannot identify the precise onset of slip, limiting its utility as a real-time feedback signal for deployed robotic controllers. Slip detection is most actionable when it occurs within tens of milliseconds of incipient contact failure; a sliding-window or causal temporal model that emits a prediction at every timestep would be far more deployable.
- **Modality imbalance persists despite gradient reweighting:** On-the-fly gradient modulation (OGM) rebalances per-modality gradient ratios during training, yet downstream F1 does not improve substantially. We attribute this to the fact that representational damage from early dominance is already encoded in the pretrained feature space before gradient reweighting can intervene; OGM operates on gradient *magnitudes* rather than on the latent geometry. Future work to meaningfully resolve modality imbalance likely requires architectural changes such as separate learning rate schedules per modality, or other explicit constraints between stream embeddings.
- **Paradoxical role of tactile:** Tactile-only models achieve modest accuracy and are consistently outperformed by RGB-only configurations, yet removing the tactile stream from full five-modal fusion produces the largest single-modality performance drop across all ablations. This paradox indicates that tactile information is not self-sufficient, but is highly complementary, it provides contact cues that other modalities cannot replace. Designing training procedures that surface and preserve this complementarity from the earliest training stages remains an open problem.
- **Broader implications for contact-rich manipulation:** As GelSight-style and gel-free tactile sensors become more accessible and commercially available, the central challenge for multimodal manipulation will shift from data acquisition to representation learning: specifically, building models that understand *how* tactile signals interact with vision and proprioception rather than treating them as yet another input channel. TEMU offers a step in this direction, but the field still lacks the pre-training corpora, standardized benchmarks, and architectural primitives that have driven progress in vision-language models.

## 9 ETHICAL CONCERNS AND CONSIDERATIONS (UNINTENTIONAL, MALICIOUS, AND DUAL-USE)

**Unintentional Harms:** Despite being designed to improve robotic safety, TEMU can cause harm through silent, undetected failure. The most immediate risk is a false negative, which is predicting a grasp as stable when it is in fact slipping. This causes the robot to continue its motion and drop or fling an object, posing a direct physical hazard in human-in-the-loop environments.

On a societal scale, systems that automate robust manipulation accelerate the displacement of manual labor in warehouses and manufacturing facilities, a consequence that is rarely attributable to any single model but accumulates across many deployments.

**Malicious Use:** TEMU is a research model for grasp stability prediction and has no direct malicious applications according to the team.

**Dual-Use:** The capabilities TEMU develops are inherently double-edged. Robust tactile grasp stability prediction transfers directly to prosthetic hand control and robotic surgical assistants, where the benefits are substantial. But failure modes in those domains are life-critical rather than merely task-critical, meaning the same overconfidence that causes a warehouse robot to drop a box could cause a surgical manipulator to mishandle anatomy.

More broadly, improved contact-rich robotic manipulation is directly applicable to autonomous systems in defense and security contexts: the same model that makes a logistics robot safer also makes an autonomous ordnance-handling platform more capable.

## 10 TEAM MEMBER CONTRIBUTIONS TO THIS REPORT

**All team members contributed equally to the poster and reports.**

**Bhaswanth Ayapilla** contributed to baselines training, MBT experiments and modality confidence score intrinsic metrics.

**Aayush Fadia** Implemented and trained the vanilla transformer. Ablated Vanilla Transformer over multiple combinations of modalities, and determined relative importance.

**Megan Lee** designed and implemented the multimodal bottleneck transformer model to work for our dataset and implemented the training pipeline.

**Parth Singh** contributed to the balanced projection, CLIP ViT+T3 encoding and training pipeline, experiments for various modifications, and to the saliency maps for the MBT.

**Ranai Srivastav** contributed to LSTM baselines, MBT architecture, ablations across moadlities, class distribution imbalance and hyperparameter tuning.

## REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Yasemin Bekiroglu, Jonna Laaksonen, Jimmy Alison Jørgensen, Ville Kyrki, and Danica Kragic. Assessing grasp stability based on learning and haptic data. *IEEE Transactions on Robotics*, 27 (3):616–629, 2011.

Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In *Conference on Robot Learning (CoRL)*, 2017.

Yevgen Chebotar, Karol Hausman, Zhe Su, Artem Molchanov, Oliver Kroemer, Gaurav S. Sukhatme, and Stefan Schaal. Bigs: Biotac grasp stability dataset. In *ICRA 2016 Workshop on Grasping and Manipulation Datasets*, 2016.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Siyuan Dong, Wenzhen Yuan, and Edward H. Adelson. Improved GelSight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2020.

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.

Jianhua Li, Siyuan Dong, and Edward Adelson. Slip detection with combined tactile and visual information. *arXiv preprint arXiv:1802.10153*, 2018.

Willow Mandil, Kiyanoush Nazari, and Amir Ghalamzan E. Action conditioned tactile prediction: Case study on slip prediction. *arXiv preprint arXiv:2205.09430*, 2024.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, abs/2008.00299, 2020. URL https://arxiv.org/abs/2008.00299.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion, 2022. URL https://arxiv.org/abs/2107.00135.

Kiyanoush Nazari, Willow Mandil, and Amir Ghalamzan E. Proactive slip control by learned slip model and trajectory adaptation. *arXiv preprint arXiv:2209.06019*, 2022.

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.

Ramprasath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Zilin Si, Zirui Zhu, Arpit Agarwal, Stuart Anderson, and Wenzhen Yuan. Grasp stability prediction with sim-to-real transfer from tactile sensing. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7809–7816, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 24043–24055, 2022.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.