Wake up for Touch! Mask-isolated Tactile Alignment Learning in MLLMs

Yoonhyung Park*, Minji Kim*, Sungwon Moon, and Jiyoung Lee†

Division of Artificial Intelligence & Software, Ewha Womans University

{pyoon0820, xxinzzi03}@ewhain.net, {sungwon268,lee.jiyoung}@ewha.ac.kr

http://mmai.ewha.ac.kr/splash/

Abstract. Touch supplies the physical grounding needed to perceive intrinsic material properties, such as friction and compliance, that vision alone often cannot resolve. Recent efforts for equipping multimodal LLMs with this tactile sense, however, expose a zero-sum trade-off: the limited parameter budget of compact models forces a choice between acquiring the new sensory modality and preserving the established vision-language reasoning. We present **Splash**, a mask-isolated tactile alignment learning framework for MLLMs. Splash quantifies the significance of each pretrained parameter, and partitions the parameter space into a dormant and critical subspace. While the frozen critical subspace acts as a stable anchor to safeguard general visual knowledge, Splash updates the isolated dormant subspace to internalize tactile alignment towards LLMs. This selective, non-destructive expansion effectively prevents catastrophic forgetting and ensures non-destructive modality expansion. Extensive experiments show that Splash effectively achieves tactile reasoning without additional inference overhead in the LLM part, demonstrating state-of-the-art performance on visuo-tactile benchmarks, including SSVTP, TVL, and TacQuad, while preserving its original general-purpose capabilities.

Keywords: Visuo-Tactile-Language Learning · Catastrophic Forgetting · Multimodal Large Language Models

1 Introduction

Humans interact with the physical world not merely by looking at it, but by touching or pressing it to judge whether a surface is slippery, compliant, or rigid. This innate ability to perceive intrinsic material properties such as friction and compliance is essential feedback for precise physical interaction. For embodied robots to achieve similar dexterity, perceiving fine-grained contact dynamics is crucial for manipulating objects. However, vision-based systems often struggle to infer such properties, which are frequently occluded or visually indistinguishable during active manipulation [4, 47].

* These authors contributed equally to this work.
† Corresponding author.

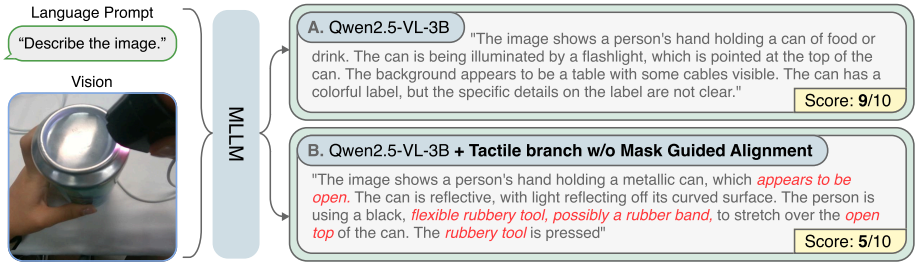


Fig. 1: Example of the catastrophic forgetting problem in tactile alignment for MLLMs (*e.g.*, TVL [21] w/ Qwen2.5-VL-3B). A small amount of tactile training set often forgets the visual sense in the base MLLM. More failure cases in Appendix.

To close this gap, recent approaches [9, 16, 29, 45] have explored learning cross-modal associations between visual appearance and tactile feedback. These approaches typically learn a shared visuo-tactile-language (VTL) representation space. While effective for discriminative tasks such as retrieval and classification, they lack the capacity for the higher-level semantic reasoning and instruction following that complex embodied decision-making demands. Moreover, their learned representations generalize poorly to novel objects with visually ambiguous textures. These observations indicate the need for models capable of reasoning over multimodal sensory inputs.

Multimodal large language models (MLLMs) [2, 8, 33] are a natural fit. By drawing on the world knowledge encoded in their LLM backbones, MLLMs support open-vocabulary reasoning over sensory inputs, allowing robots not only to perceive but also to reason about the physical consequences of their actions. Most existing MLLMs, however, rely on computationally intensive backbones (*e.g.*, LLaMA-7B [41]), which limit their practicality for resource-constrained edge robots [43]. This motivates compact MLLMs capable of tactile reasoning under real-world deployment constraints [11].

The challenge is most acute for small MLLMs (sMLLMs, *e.g.*, $\leq$ 3B-scale), which are the ones actually desirable for edge deployment. Their limited capacity means that incorporating a new touch modality interferes with existing vision-language representations. As Fig. 1 illustrates, optimizing such a model for tactile signals distorts its pretrained visual features, causing catastrophic forgetting and degraded vision-language (VL) reasoning performance. The result is a fundamental trade-off between expanding sensory capabilities and preserving the pretrained reasoning abilities of sMLLMs.

In this paper, we propose SPLASH, a novel framework that aligns tactile signals into sMLLMs while mitigating catastrophic forgetting [28, 39, 49]. The key idea is to confine all tactile adaptation to a dormant (*i.e.*, less important) parameter subspace while keeping the critical parameters frozen. We estimate parameter importance with a VL-relative metric computed from weight and activation statistics, then restrict gradient updates to the dormant subspace. Since

frozen critical parameters are stable vision-language anchors, our selective optimization approach preserves pretrained reasoning capabilities while enabling the model to incorporate tactile information. By isolating and reactivating internal weights, SPLASH maintains the original MLLM scale without introducing external modules like adapters, aside from a negligible tactile front-end. Moreover, SPLASH supports a unified, single-stage training, without multi-stage alignment and fine-tuning procedures [21, 44]. Extensive experiments on visuo-tactile-language (VTL) benchmarks, including SSVTP [26], TVL [21], and TacQuad [16], validate the effectiveness of SPLASH. Notably, SPLASH-3B achieves state-of-the-art (SOTA) performance, outperforming previous 7B-scale methods [21,44]. In addition, we present the first comprehensive performance analysis of preservation in VTL alignment across complex general-purpose VL benchmarks. The results demonstrate that tactile alignment with SPLASH does not compromise the broad-spectrum intelligence of MLLM backbones.

We summarize our contributions as follows:

1. We introduce SPLASH, a parameter-isolated tactile adaptation framework that integrates tactile sensing into sMLLMs with zero additional inference overhead, mitigating the vision-related catastrophic forgetting.
2. By leveraging frozen critical parameters as stable structural anchors, SPLASH enables single-stage unified training that optimizes tactile front-end and the dormant LLM subspace concurrently.
3. Across small-scale MLLMs including Qwen2.5-VL-3B and InternVL-1B, SPLASH achieves SOTA performance on three VTL benchmarks while safeguarding the models' pretrained general-purpose capabilities.

2 Related Work

2.1 Visuo-Tactile-Language (VTL) Alignment

The integration of vision and tactile sensing has long been central to robotic perception, chiefly as a way to resolve the ambiguities of any single modality. Early work established the value of synergistic visuo-tactile integration across a spectrum of robotic tasks such as force estimation [6, 46], cross-modal synthesis [45, 47], and robotic manipulation [4, 22]. With the rise of MLLMs [2, 8, 33], the attention has shifted toward VTL alignment, where the tactile modality is aligned with a vision-language latent space, typically via contrastive learning objectives [9, 44, 45]. Subsequent efforts push scalability further by constructing large-scale VTL datasets [9, 21] and by learning cross-sensor visuo-tactile representations [16,17,44] that generalize across heterogeneous tactile sensors. Despite these advances, how to integrate tactile sensing into compact multimodal models remains largely underexplored. In this work, we address this gap directly, studying how the tactile modality can be folded into small MLLMs without sacrificing their pretrained abilities.

2.2 Pruning-based Multi-Task Learning

Our work shares the underlying principle of pruning-based multi-task learning [18, 38], where disjoint parameter subsets are allocated to different tasks to mitigate cross-task interference. Recent findings in LLMs [19, 24, 27, 37] reinforce this view: updating only a small fraction of parameters can achieve competitive performance compared to full finetuning while effectively curbing catastrophic forgetting. These results imply that model capacity can be selectively exploited without disturbing the entire parameter space. However, these works predominantly focus on intra-modality adaptation, such as continual learning [28, 31, 39], primarily aiming to pack multiple tasks into one model while minimizing interference among them. We instead extend this principle from task-level to modality-level parameter isolation, addressing the severe representation gap encountered in new tactile sensory alignment.

3 Splash

3.1 Motivation and Problem Formulation

MLLMs [8, 30, 41, 50] typically consist of a vision encoder, a modality adapter, and an LLM that reasons over the fused multimodal input. Of these, the LLM's scale largely governs reasoning capacity and generalization. Enlarging the LLM improves performance on complex tasks and zero-shot generalization, but at a steep computational cost. This escalation in model size presents a critical bottleneck for robotic applications, where real-time inference and limited onboard compute are required [11, 14, 51]. Worse, naively integrating a new sensory modality risks catastrophic forgetting: tactile instruction tuning that updates the LLM parameters can overwrite the weight configurations encoding pretrained visual knowledge, even when only a small amount of tactile data is used.

To address these challenges, we propose Mask-guided tactile alignment learning, Splash. Given a natural (*i.e.*, RGB) image X_R, a tactile contact image X_C, and a text prompt X_T, the goal is to generate a tactile description Y in natural language form:

$$Y = \texttt{LLM}_\theta(\mathcal{F}_\phi(X_R), \mathcal{F}_\psi(X_C), X_T), \tag{1}$$

where θ are the language-model parameters. We denote the visual front-end, consisting of a vision encoder and modality adapter, as $\mathcal{F}$, with the natural-image and tactile front-ends parameterized by ϕ and ψ, respectively. We initialize $\texttt{LLM}_\theta$ and $\mathcal{F}_\phi$ using weights from a pretrained sMLLM [3] to inherit its general-purpose reasoning and visual understanding. For the tactile front-end $\mathcal{F}_\psi$, we employ an ImageNet-pretrained [12] vision transformer (ViT) [13] to extract geometric features from contact images. During training, we freeze the visual parameters ϕ and update the tactile parameters ψ together with a subset of the LLM parameters θ, determined by a sparsity ratio τ. Next, we describe the key components of our proposed framework in detail.

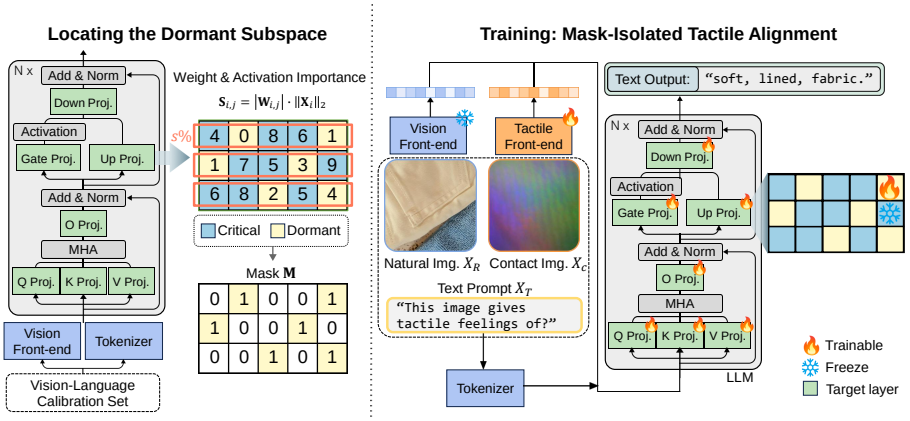


Fig. 2: The overall framework of SPLASH. To mitigate the catastrophic forgetting in the visual aspect, we train a dormant subspace in LLM for tactile alignment in sMLLMs. We note that the tactile front-end is also updated in a unified training stage.

3.2 Locating the Dormant Subspace

Given a pretrained MLLM (*e.g.*, QwenVL-2.5 [3], InternVL [8]), SPLASH first identifies a dormant parameter subspace that contributes minimally to pretrained VL tasks. This builds on the *natural sparsity* hypothesis [18], which holds that only a subset of parameters is critical for maintaining model function. We exploit such redundancy by repurposing these underutilized weights, referred to as the **dormant subspace**, as trainable capacity for the tactile modality.

To identify dormant parameters, we use a visual-relative importance metric inspired by Wanda [40]. With a small calibration set drawn from VL data (*e.g.*, 128 samples from CC12M [5]), we score the importance of each weight $\mathbf{W}_{i,j}$ connecting the i-th input feature to the j-th output neuron in the LLM's linear layers. The resulting scores reflect each parameter's contribution to pretrained VL behavior. Formally, the importance score $\mathbf{S}_{i,j}$ is defined as:

$$\mathbf{S}_{i,j} = |\mathbf{W}_{i,j}| \cdot \|\mathbf{x}_i\|_2, \tag{2}$$

where $|\cdot|$ denotes the weight magnitude and $\|\mathbf{x}_i\|_2$ denotes the ℓ_2-norm of the corresponding input activation. Given a sparsity ratio $s\%$, we set a threshold $\tau_j(s)$ such that $s\%$ of the scores in $\{\mathbf{S}_{i,j}\}_{i=1}^{d_{\text{in}}}$ fall below it, and form the binary mask within each layer:

$$\mathbf{M}_{i,j} = \mathbb{1}\Big(\mathbf{S}_{i,j} < \tau_j(s)\Big), \tag{3}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. For example, when $s = 0\%$, all parameters are treated as critical and remain frozen, whereas $s = 100\%$ corresponds to full-parameter tuning of the LLM. We additionally set $\mathbf{M}_{i,j} = 0$ for the first and last transformer blocks, preserving input grounding and high-level semantic

representations. It thereby ensures training stability and prevents representation collapse during tactile adaptation [1, 40]. Consequently, the resulting mask cleanly partitions the LLM into frozen critical weights and trainable dormant weights, enabling tactile alignment while protecting the pretrained VL capabilities of the model.

3.3 Mask-isolated Tactile Alignment

Existing VTL alignment methods [16, 21] typically follow a two-stage training scheme. First, the tactile encoder is aligned with either the vision-language [21] or the vision representation space [16]. Second, the LLM is adapted using parameter-efficient finetuning (PEFT) (*e.g.*, LoRA [23], $(IA)^3$ [32]) for multimodal reasoning. SPLASH instead enables a unified, single-stage training process by leveraging the parameter partition from Sec. 3.2. Specifically, we optimize the tactile frontend $\mathcal{F}_\psi$ and the dormant LLM parameters while holding the critical parameters frozen. The training objective $\mathcal{L}$ maximizes the likelihood of the autoregressively generated tactile description as follows:

$$\mathcal{L} = -\sum_{n=1}^{N} \log P(y_n | Y_{<n}, X_R, X_C, X_T), \tag{4}$$

where N is the number of tokens in the output sequence Y. We apply the binary mask $\mathbf{M}$ to the LLM backbone parameters, while the tactile parameters ψ remain fully trainable. Let η denote the learning rate. The parameter updates are defined as:

$$\theta^* \leftarrow \theta - \eta \left(\mathbf{M} \odot \nabla_\theta \mathcal{L}\right), \tag{5}$$

$$\psi \leftarrow \psi - \eta \nabla_\psi \mathcal{L}, \tag{6}$$

where $\odot$ denotes Hadamard multiplication. Eq. (5) gives the critical parameters (*i.e.*, $\mathbf{M}_{i,j} = 0$) zero gradient and dormant parameters (*i.e.*, $\mathbf{M}_{i,j} = 1$) for gradient update, while Eq. (6) allows the tactile branch to train fully.

Given that we preserve the critical vision-language weights, these frozen parameters work as stable anchors that maintain the integrity of the multimodal manifold during adaptation. Crucially, unlike additive PEFT methods such as LoRA ($\theta + \Delta\theta$), which can perturb the entire parameter manifold, SPLASH employs strict structural isolation by confining updates exclusively to the dormant subspace. This sparse replacement mechanism ensures that the pretrained VL reasoning pathways remain fundamentally untouched, effectively preventing catastrophic forgetting without any additional inference latency. Therefore, this design eliminates the need for a separate alignment stage and simplifies the training pipeline, significantly reducing the overall training cost, and cohesively integrating the tactile into the reasoning process of MLLMs.

4 Experiments

4.1 Datasets

Visuo-Tactile-Language (VTL). For VTL alignment learning, we train SPLASH on DIGIT-sensor-based datasets, SSVTP [26] and TVL [21]. SSVTP contains 4.5K aligned visuo-tactile pairs, whereas TVL contains 44K pairs annotated with natural language descriptions of tactile properties. We use the official train/test split and further divide the training set into train/validation at a 9:1 ratio. To assess robustness under distribution shift, we further evaluate on the DIGIT-sensor subset of TacQuad [16] (72K tactile samples), collected via handheld interaction to capture diverse real-world contact dynamics of humans. This dataset includes a detailed linguistic description, initially generated by GPT-4o [25] and manually revised. By leveraging this richly annotated dataset, our evaluation enables a more natural language-based assessment beyond the keyword-based evaluation.
General-purpose Vision-Language (VL). To assess retention of pretrained VL capabilities after tactile adaptation, we evaluate models on standard vision-language benchmarks, including MMMU [48] for expert-level multi-disciplinary reasoning, MathVista [36] for complex mathematical logic within visual context, MME [20] for broad perception and cognitive stability, and MMBench [34] for assessment on bilingual context. These benchmarks evaluate diverse multimodal reasoning and perception capabilities, allowing us to assess whether tactile adaptation preserves the pretrained VL capabilities of the model.

4.2 Evaluation Protocol

Following the evaluation protocol of TVL [21], we evaluate tactile semantic generation on SSVTP, TVL and TacQuad using an LLM-based evaluator (*i.e.*, LLM-as-a-judge). Since TacQuad consists of trajectory-level tactile and visual image sequences paired with one free-form textual description, we use the middle frame of each contact trajectory as the representative observation for evaluation. Specifically, we adopt GPT-4o [25] as a text-only judge to score the semantic similarity between the generated tactile descriptions and ground-truth tactile descriptions. The score ranges from 1 to 10, where higher scores indicate better semantic alignment. The score is computed as the average GPT-4o score across all samples (see Sec. B in Appendix for the full evaluation prompt). To complement GPT-4o semantic evaluation and reduce potential stylistic bias from LLM-based judging, we additionally report objective tactile metrics including keyword-based F1 score and Top-5 Accuracy. F1 is computed based on tactile keyword overlap, while Top-5 Accuracy measures whether at least one predicted tactile attribute matches the ground-truth set.

4.3 Implementation Details

To validate SPLASH, we employ two high-performance, open-source sMLLMs: Qwen2.5-VL 3B [3] and InternVL2.5-1B [7]. These models represent the current

state-of-the-art in compact multimodal intelligence, offering an ideal balance between sophisticated reasoning and deployment efficiency. To minimize the computational footprint, we select a tactile encoder as a ViT-Tiny [13] with 16×16 patches, initialized from ImageNet-pretrained [12] weights. This encoder has only 5.8M parameters. A two-layer MLP projector with LayerNorm and GELU activation maps tactile features into the language embedding space. Both natural and tactile contact images are resized to 224×224. Following the Qwen2.5-VL formulation, tactile, visual, and textual tokens are concatenated into a single autoregressive sequence. Training is performed using the AdamW optimizer [35] with weight decay set to 0.0. For Splash-3B, we train for 1 epoch with an initial learning rate of 2×10^{-5}, cosine decay scheduling and a warmup ratio of 0.05. For Splash-1B, we train for 2 epochs using the same optimization setup. All experiments are conducted on two NVIDIA RTX PRO 6000 Blackwell GPUs. For mask-guided fine-tuning, the sparsity ratio is set to $s = 60\%$. For LoRA [23] fine-tuning, we use rank $r = 8$ with $\alpha = 16$. During inference, we use the following instruction *"List 2-3 tactile attributes of the object shown in the image. Use single-word descriptors only. Output a comma-separated list with no additional text."*, to generate tactile descriptions.

4.4 Baselines

We compare Splash with existing VTL approaches that extend LLMs including TVL-LLaMA [21] and UniTouch (Touch-LLM) [44]. Both methods employ 7B-parameter LLaMA models as the backbone, with UniTouch using LLaMA [41] and TVL-LLaMA using LLaMA-2 [42], and initialize the vision and tactile encoders using OpenCLIP [10]. Among multiple tactile encoder scales (*e.g.*, Tiny, Small, Base), we follow the ViT-Tiny setting. UniTouch adopts a larger ViT-L/14 encoder in its original design, and we follow this default configuration in our experiments. Both approaches [21, 44] align tactile features with the language model and adapt the LLM using LoRA [23]. By contrast, our Splash adopts modern vision-language models (VLMs), including Qwen2.5-VL-3B and InternVL2.5-1B, as the backbone. Since LLaMA-7B is a language-only backbone, it requires additional visual instruction tuning to perform zero-shot multimodal tasks. Therefore, to ensure a rigorous and fair comparison, we established a TVL-style LoRA baseline using the same Qwen2.5-VL-3B backbone adopted in Splash. This comparison isolates the effect of the adaptation strategy (LoRA vs. mask-guided fine-tuning) under identical initialization and training conditions. We further compare against $(IA)^3$ [32] to assess a different PEFT paradigm that introduces learned scaling vectors instead of additive weight updates.

4.5 Performance Comparison on VTL Tasks

Tab. 1 compares performance on VTL tasks across three benchmarks. We include zero-shot performance of pretrained VLMs without tactile adaptation as baselines. Without tactile alignment, the zero-shot InternVL-1B and Qwen2.5-VL-3B

Table 1: Overall performance comparison using LLM Judge to record scores in 10-point scale on three VTL benchmarks, including SSVTP [26], TVL [21], and TacQuad [16]. **The best score** is bold, and the second is underlined.

Method	Backbone	SSVTP	TVL	TacQuad	Avg
Zero-shot	InternVL2.5-1B	3.61	3.54	3.45	3.53
Zero-shot	Qwen2.5-VL-3B	3.74	3.40	3.57	3.57
UniTouch [44]	LLaMA1-7B	3.73	4.04	3.44	3.74
TVL [21]	LLaMA2-7B	5.27	4.38	3.29	4.31
TVL [21]	Qwen2.5-VL-3B	4.98	4.29	4.22	4.50
TVL-$(IA)^3$ [21, 32]	Qwen2.5-VL-3B	5.17	4.26	3.84	4.42
SPLASH-1B	InternVL2.5-1B	**5.69**	4.34	**5.02**	**5.01**
SPLASH-3B	Qwen2.5-VL-3B	5.48	**4.39**	4.86	4.91

models achieve average scores of only 3.53 and 3.57, respectively, indicating limited capability in reasoning about tactile properties. Surprisingly, SPLASH-1B and -3B significantly outperform UniTouch, which uses a larger LLaMA1-7B backbone yet achieves only 3.74 on average. This highlights the effectiveness of the proposed mask-isolated adaptation even with smaller MLLMs.

Under the same Qwen2.5-VL-3B, SPLASH-3B achieves the best tactile-semantic performance among methods using the same backbone, reaching an average score of 4.91. In comparison, the strongest baseline, TVL [21], with LoRA-based adaptation on Qwen2.5-VL-3B, achieves 4.50 on average. In particular, SPLASH-3B outperforms TVL 5.48 compared to 4.98 on SSVTP and 4.86 to 4.22 on TacQuad, suggesting strong visuo-tactile alignment and robustness under tactile distribution shifts. While $(IA)^3$ shows competitive performance on relatively smaller VTL datasets like SSVTP and TVL, it suffers a significant performance drop on the unseen TacQuad, scoring only 3.84 (vs. 4.86 for SPLASH). This suggests that $(IA)^3$ lacks the representational capacity required to generalize to real-world tactile dynamics.

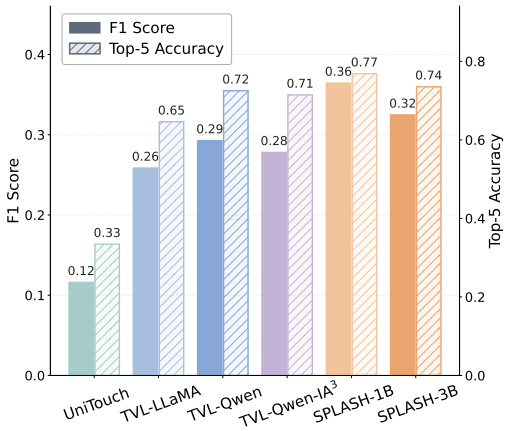


Fig. 3: Quantitative comparison with objective metrics on SSVTP [26], TVL [21], and TacQuad [16]. We report the averaged F1 score and Top-5 accuracy. SPLASH demonstrates superior performance in both F1 and Top-5 Accuracy.

To verify that the effectiveness of our approach is not limited to a specific scale, we further employ SPLASH with a smaller InternVL-1B backbone. Despite

Table 2: Overall performance comparison on five VL benchmarks, such as MMMUval [48], MathVista [36], MMEsum [20], MMBench-EN [34], and MMBench-CN [34]. **The best score** is bold, and the second is underlined.

Method	Backbone	MMMUval	MathVista	MMEsum	MMBench-EN	MMBench-CN
Zero-shot	InternVL2.5-1B	40.9	43.2	1950	70.7	66.3
Zero-shot	Qwen2.5-VL-3B	53.1	62.3	2157	79.1	78.1
UniTouch [44]	LLaMA1-7B	26.6	9.0	27	8.5	8.6
TVL [21]	LLaMA2-7B	26.0	11.6	582	8.6	5.5
TVL [21]	Qwen2.5-VL-3B	50.0	52.8	2168	76.8	76.5
TVL-$(IA)^3$ [21, 32]	Qwen2.5-VL-3B	51.3	63.0	**2216**	**78.4**	**78.9**
Splash-1B	InternVL2.5-1B	37.3	41.0	1786	70.2	60.4
Splash-3B	Qwen2.5-VL-3B	**55.3**	**65.3**	2155	78.0	76.9

its significantly smaller parameter size, Splash-1B achieves competitive tactile-semantic performance, reaching an average score of 5.01. These results suggest that the proposed mask-guided adaptation remains effective even for compact VLM backbones.

Beyond LLM-based evaluation, we quantitatively assess tactile attribute prediction with objective metrics. As shown in Fig. 3, Splash consistently outperforms existing baselines in both F1 score and Top-5 Accuracy, indicating more precise tactile attribute generation and stronger robustness across diverse tactile vocabularies.

4.6 Performance Comparison for Preservation of VL Capability

Tab. 2 evaluates VL capability retention after tactile adaptation on standard VL benchmarks. Splash-3B consistently outperforms the TVL baselines across all VL benchmarks. In particular, Splash improves MMMUval from 50.0 to 55.3, MathVista from 52.8 to 65.3, and maintains comparable performance on MMEsum. Furthermore, Splash achieves higher scores on both MMBench-EN and MMBench-CN compared to TVL. Compared with UniTouch [44], which relies on a larger LLaMA-7B backbone, Splash achieves substantially stronger VL reasoning performance across all benchmarks.

We note that TVL-$(IA)^3$ [21, 32] attains the highest scores on three benchmarks (*i.e.*, MMEsum, MMBench-EN, and MMBench-CN), as its scaling-based adaptation perturbs the pretrained weights only minimally. This apparent advantage, however, comes at the expense of tactile learning: as shown in Tab. 1, TVL-$(IA)^3$ fails to generalize on TacQuad (3.84 vs. 4.86 for Splash-3B), revealing that its limited plasticity preserves VL ability only by under-fitting the tactile modality. In contrast, Splash attains the best scores on MMMUval and MathVista while remaining competitive on the remaining benchmarks, demonstrating that mask-guided isolation uniquely balances tactile acquisition with VL preservation – rather than trading one for the other. Notably, the performance of Splash remains comparable to the zero-shot Qwen2.5-VL-3B baseline, which achieves 53.1 on MMMUval and 62.3 on MathVista, confirming that tactile

Table 3: The impact of sparsity ratio in Splash-3B. **The best score** is bold, and the second is underlined. Based on the results, we establish 60% as our base sparsity ratio (s=0.6).

Sparsity	Visuo-Tactile-Language				Vision-Language				
	SSVTP	TVL	TacQuad	Avg	MMMUval	MathVista	MMEsum	MMBench-EN	MMBench-CN
100%	5.02	4.36	4.81	4.73	22.0	38.4	982	10.4	28.3
80%	5.37	4.33	4.85	4.85	52.6	57.9	2070	56.0	69.5
60%	5.48	**4.39**	**4.86**	**4.91**	**55.3**	65.3	2155	78.0	76.9
50%	5.10	4.32	4.74	4.72	54.0	**65.8**	2167	78.3	77.7
40%	**5.60**	4.36	4.74	4.90	52.6	65.7	2204	**78.6**	78.6
30%	5.30	4.27	4.79	4.79	52.0	65.4	**2210**	78.0	**78.8**

adaptation does not degrade the pretrained VL capability. We hypothesize that selectively updating VL-dormant subspaces acts as a form of pruning-induced regularization, which stabilizes the pretrained reasoning manifold during tactile adaptation [15].

4.7 Ablation Study

Sparsity ratio. Tab. 3 presents an ablation study on the sparsity ratio used for mask-guided adaptation and its impact on both VTL and VL performance. We observe that moderate sparsity levels provide the best trade-off between tactile-semantic performance and VL capability retention. In particular, 60% sparsity achieves the highest tactile-semantic average score of 4.91 while simultaneously obtaining the best VL reasoning performance on MMMUval and competitive results on MathVista and the MMBench benchmarks. When the sparsity ratio becomes too large (*e.g.*, 100%), excessive parameter updates severely degrade VL reasoning performance, indicating catastrophic forgetting of pretrained knowledge. Conversely, when the sparsity ratio becomes too small (*e.g.*, 30%), excessive parameter freezing slightly limits tactile adaptation, leading to reduced tactile-semantic performance while VL capability remains largely preserved. These results highlight the importance of allocating an appropriate dormant subspace that balances tactile learning capacity with preservation of pretrained VL capabilities. The analysis on the sparsity ratio for InternVL-1B is further reported in Sec. A.2 in Appendix.

Calibration data. Tab. 4 evaluates the effectiveness of different calibration datasets used to construct the dormant subspace mask for Splash-3B. We first ablate the calibration sample size. The proposed method remains highly stable across 64, 128, and 256 samples, showing only marginal variations across benchmarks. While all configurations achieve competitive performance, we select 128 samples as the default setting since it provides the optimal balance among VL reasoning performance, stability, and calibration efficiency. In addition, using either CC3M or TVL for the calibration dataset yields comparable performance on both VTL and VL benchmarks (Tab. 4(b)). The VTL average score changes only slightly from 4.91 to 4.75, while the VL benchmark results remain largely stable.

Table 4: Performance evaluation on calibration data choices, configurations, and sample sizes on SPLASH-3B. SPLASH demonstrates robust performance across varying setups and remains highly competitive even with unpaired noise data, indicating that it captures intrinsic architectural redundancies rather than overfitting to specific semantic alignments.

Calibration Setup	Visuo-Tactile-Language				Vision-Language				
	SSVTP	TVL	TacQuad	Avg	MMMUval	MathVista	MMEsum	MMBench-EN	MMBench-CN
(a) Sample Sizes (Data = CC3M Paired)									
64 Paired	5.15	4.30	4.87	4.77	**56.0**	64.9	**2168**	77.7	**77.4**
128 Paired (Ours)	**5.48**	**4.39**	4.86	**4.91**	55.3	65.3	2155	**78.0**	76.9
256 Paired	5.28	4.35	4.86	4.83	**56.0**	**65.6**	2146	77.7	77.2
(b) Data Choices & Configurations (Sample Size = 128)									
CC3M Paired (Ours)	**5.48**	**4.39**	4.86	**4.91**	55.3	**65.3**	2155	**78.0**	76.9
TVL Paired	4.98	4.35	**4.93**	4.75	**58.0**	62.9	2156	76.5	76.8
Unpaired (Noise)	5.45	4.31	4.89	4.88	55.3	64.9	**2166**	77.7	**77.4**

Table 5: Comparison using the same tactile-pretrained frontend initialization adopted in TVL.

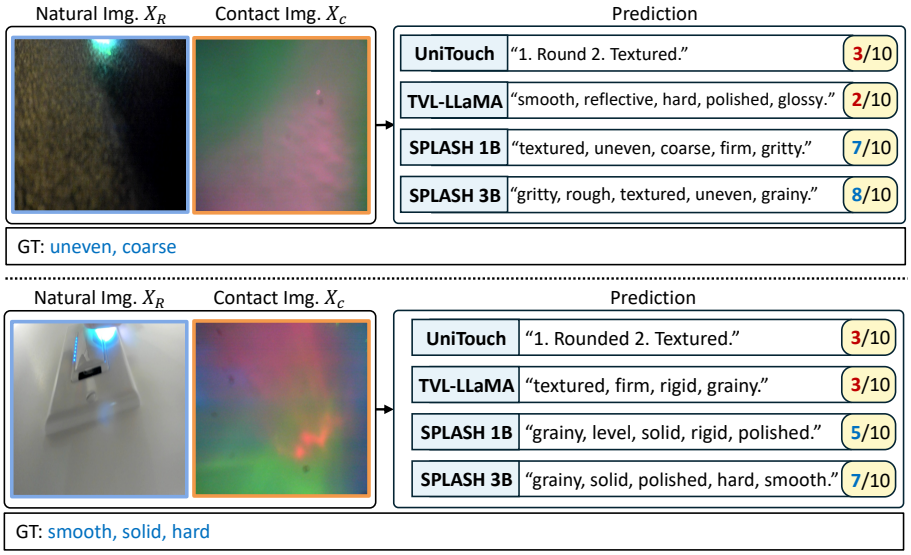
Method	Visuo-Tactile-Language				Vision-Language				
	SSVTP	TVL	TacQuad	Avg	MMMUval	MathVista	MMEsum	MMBench-EN	MMBench-CN
TVL(Qwen2.5-VL-3B)	4.98	4.29	4.22	4.50	50.0	52.8	2168	76.8	76.5
SPLASH-3B	**5.10**	**4.35**	**4.88**	**4.78**	**56.0**	**65.3**	**2195**	**78.7**	**76.7**

These observations suggest that the proposed visual-relative importance metric is robust to the choice of calibration corpus. Most surprisingly, we further assess an extreme setup where vision-language semantically calibrated data is unavailable (*i.e.*, using unpaired random noise images with semantically meaningless dummy text prompts). The performance on the 'unpaired' setting in Tab. 4(b) shows this calibration-free variant still achieves highly competitive performance. Consequently, SPLASH successfully captures the intrinsic architectural redundancies of modern MLLMs, rather than merely overfitting to dataset-specific activation patterns.

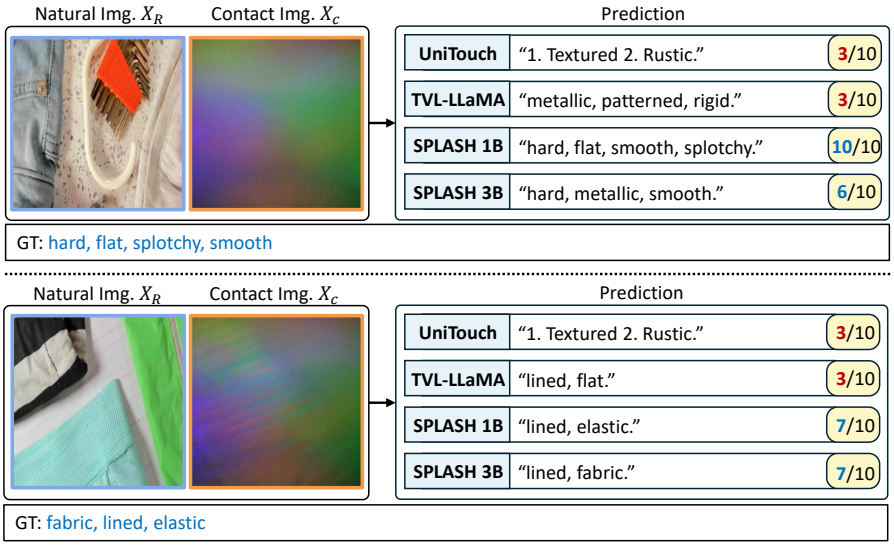
Tactile frontend initialization. To isolate the effect of tactile frontend initialization, we additionally compare SPLASH-3B and TVL using the same tactile-pretrained encoder initialization adopted in TVL. In Tab. 5, SPLASH consistently outperforms TVL across both VTL and VL benchmarks, indicating that the performance gain mainly stems from the proposed mask-guided adaptation strategy rather than the frontend initialization itself.

4.8 Qualitative Results

Figs. 4 and 5 present qualitative comparisons across the VTL benchmarks, including TVL, SSVTP, and TacQuad. Given paired RGB and tactile contact images, models are asked to generate tactile attribute descriptions. Across all benchmarks, SPLASH produces closely matched descriptions of the ground-truth



(a) TVL



(b) SSVTP

Fig. 4: Qualitative comparisons on TVL [21] and SSVTP [26]. SPLASH demonstrates the robustness with higher accuracy score than TVL-LLaMA7B [21] and UniTouch [44]. Especially, SPLASH-1B highlights the best scores in examples on SSVTP, even at the 1B-parameter scale.

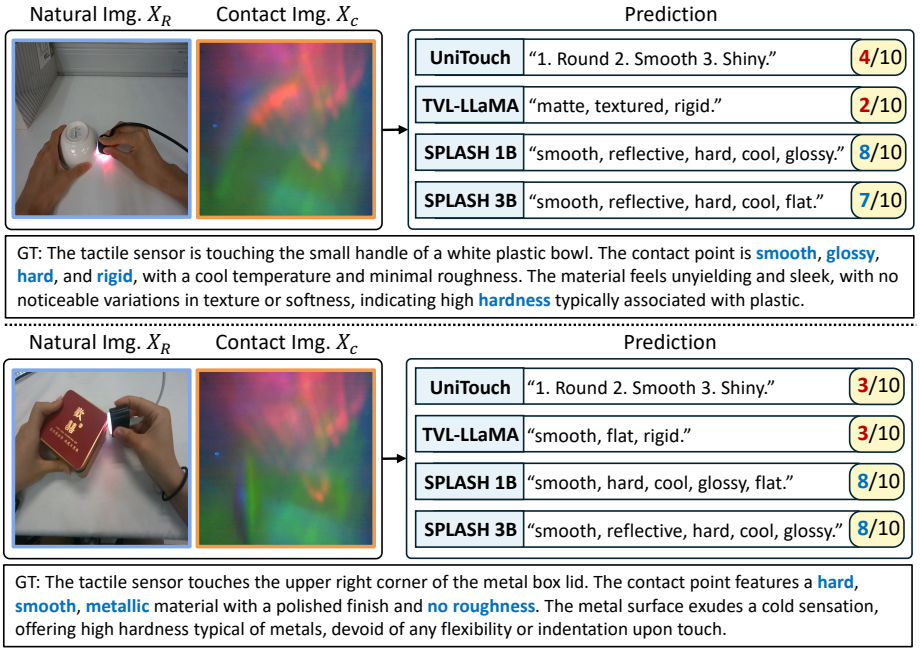


Fig. 5: Qualitative comparisons on TacQuad [16]. Both SPLASH-1B and -3B generate more diverse predictions than baselines, achieving higher accuracy scores by an LLM judge.

tactile properties. TVL-LLaMA and UniTouch often generate visually plausible but tactually inconsistent attributes (*e.g.*, smooth or reflective for coarse surfaces) or generic attributes unrelated to the underlying tactile properties. These mismatches lead to lower semantic similarity scores in the GPT-4o evaluation, whereas our models consistently achieve higher scores across the examples. Overall, the results suggest that baselines rely more heavily on visual appearance cues when tactile grounding is weak, whereas SPLASH more effectively integrates tactile observations to infer physically consistent surface properties. More qualitative examples are provided in Sec. A.8 in Appendix.

4.9 Limitation and Future Work

While SPLASH demonstrates a robust capability for non-destructive tactile adaptation, the critical subspace is static after the offline step to split parameters. In practical robotic manipulation scenarios, however, activation patterns may dynamically change depending on interaction state, contact condition, or task progression. Thus, a static dormant subspace may not fully capture temporally varying parameter utilization during long-horizon visuo-tactile reasoning. Future work explores dynamically adaptive masking strategies that update dormant parameter allocation according to task-dependent activation statistics during em-

bodied interaction. Moreover, our current experiments focus on a single DIGIT tactile sensor setup, and cross-sensor generalization remains an important direction for future investigation. Nevertheless, as SPLASH is designed in a frontend-agnostic manner, it can be naturally extended to incorporate alternative tactile encoders.

5 Conclusion

We present **SPLASH**, a novel mask-isolated tactile alignment framework that expands the sensory capability of small MLLMs without catastrophic forgetting. SPLASH identifies and isolates a dormant parameter via a visual-relative importance metric. During training, only the dormant subspace is updated simultaneously with the tactile front-end, while the critical parameters responsible for foundational VL reasoning stay anchored. Extensive experiments across diverse VTL and VL benchmarks demonstrate that SPLASH effectively "wake up" underutilized parameters to internalize complex contact dynamics, achieving state-of-the-art tactile reasoning while fully preserving general-purpose ability. With its unified single-stage training and zero added inference cost, we believe SPLASH offers a practical path toward high-level tactile intelligence for resource-constrained edge robots.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-16065706).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Calandra, R., Owens, A., Jayaraman, D., Lin, J., Yuan, W., Malik, J., Adelson, E.H., Levine, S.: More Than a Feeling: Learning to grasp and regrasp using vision and touch. IEEE RAL (2018)
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)

6. Chelly, E., Cherubini, A., Fraisse, P., Ben Amar, F., Khoramshahi, M.: Tactile-based force estimation for interaction control with robot fingers. arXiv preprint arXiv:2411.13335 (2025)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024)
9. Cheng, N., Guan, C., Gao, J., Wang, W., Li, Y., Meng, F., Zhou, J., Fang, B., Xu, J., Han, W.: Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. Information Fusion (2025)
10. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023)
11. Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, X., Hu, Y., Lin, X., Zhang, B., et al.: MobileVLM V2: Faster and stronger baseline for vision language model. arXiv preprint arXiv:2402.03766 (2024)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
14. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: PaLM-E: An embodied multimodal language model. In: ICML (2023)
15. Evci, U., Gale, T., Menick, J., Castro, P.S., Elsen, E.: Rigging the lottery: Making all tickets winners. In: ICML. PMLR (2020)
16. Feng, R., Hu, J., Xia, W., Gao, T., Shen, A., Sun, Y., Fang, B., Hu, D.: AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In: ICLR (2025)
17. Feng, R., Zhou, Y., Mei, S., Zhou, D., Wang, P., Cui, S., Fang, B., Yao, G., Hu, D.: AnyTouch 2: General optical tactile representation learning for dynamic tactile perception. In: ICLR (2026)
18. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: ICLR (2019)
19. Frantar, E., Alistarh, D.: SparseGPT: Massive language models can be accurately pruned in one-shot. In: ICML (2023)
20. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
21. Fu, L., Datta, G., Huang, H., Panitch, W.C.H., Drake, J., Ortiz, J., Mukadam, M., Lambeta, M., Calandra, R., Goldberg, K.: A touch, vision, and language dataset for multimodal alignment. In: ICML (2024)
22. Heng, L., Geng, H., Zhang, K., Abbeel, P., Malik, J.: ViTacFormer: Learning cross-modal representation for visuo-tactile dexterous manipulation. arXiv preprint arXiv:2506.15953 (2025)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)

24. Huang, W., Cheng, A., Wang, Y.: Mitigating catastrophic forgetting in large language models with forgetting-aware pruning. In: EMNLP (2025)
25. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
26. Kerr, J., Huang, H., Wilcox, A., Hoque, R., Ichnowski, J., Calandra, R., Goldberg, K.: Self-supervised visuo-tactile pretraining to locate and follow garment features. arXiv preprint arXiv:2209.13042 (2022)
27. Khaki, S., Li, X., Guo, J., Zhu, L., Plataniotis, K.N., Yazdanbakhsh, A., Keutzer, K., Han, S., Liu, Z.: SparseLoRA: Accelerating llm fine-tuning with contextual sparsity. In: ICML (2025)
28. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. USA (2017)
29. Lei, W., Ge, Y., Yi, K., Zhang, J., Gao, D., Sun, D., Ge, Y., Shan, Y., Shou, M.Z.: ViT-Lens: Towards omni-modal representations. In: CVPR (2024)
30. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
31. Li, Z., Hoiem, D.: Learning without forgetting. IEEE TPAMI (2017)
32. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.A.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In: NeurIPS (2022)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: CVPR (2023)
34. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: MMBench: Is your multi-modal model an all-around player? In: ECCV (2024)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In: ICLR (2024)
37. Ma, X., Fang, G., Wang, X.: LLM-Pruner: On the structural pruning of large language models. In: NeurIPS (2023)
38. Mallya, A., Lazebnik, S.: PackNet: Adding multiple tasks to a single network by iterative pruning. In: CVPR (2018)
39. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation. Elsevier (1989)
40. Sun, M., Liu, Z., Bair, A., Kolter, J.Z.: A simple and effective pruning approach for large language models. In: ICLR (2024)
41. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
42. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
43. Williams, J., Gupta, K.D., George, R., Sarkar, M.: Lite VLA: Efficient vision-language-action control on cpu-bound edge robots. arXiv preprint arXiv:2511.05642 (2025)

44. Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., Wong, A.: Binding touch to everything: Learning unified multimodal tactile representations. In: CVPR (2024)
45. Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., Owens, A.: Touch and Go: Learning from human-collected vision and touch. In: NeurIPS (2022)
46. Yuan, W., Dong, S., Adelson, E.H.: GelSight: High-resolution robot tactile sensors for estimating geometry and force. Sensors (2017)
47. Yuan, W., Wang, S., Dong, S., Adelson, E.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: CVPR (2017)
48. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024)
49. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language models (2023)
50. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
51. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)