Learning Visual Contact Representations from Triboelectric Signals for Tactile-Free Robot Manipulation

Anonymous authors

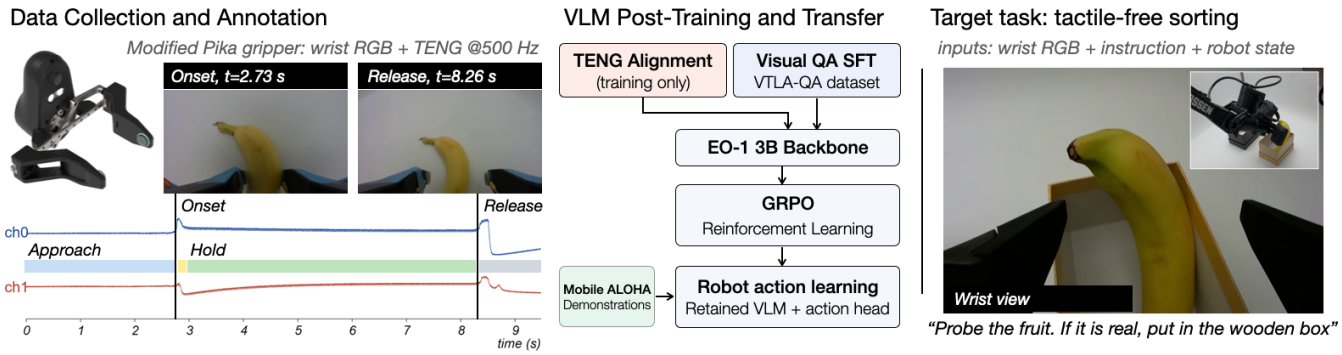

Fig. 1. Overview of the framework. **Left:** A modified Pika gripper records paired wrist RGB and TENG signals with interaction annotations. **Middle:** Joint QA SFT and alignment adapt EO-1, followed by visual QA GRPO and action learning. **Right:** A recorded Mobile ALOHA demonstration illustrates sorting. Deployment uses wrist RGB, instructions, and robot state.

***Abstract*— Tactile recordings may provide supervision for visual robot learning even when tactile sensing is unavailable at deployment. We investigate whether raw triboelectric nanogenerator (TENG) signals improve a vision-language model (VLM) beyond supervision from Question Answering (QA) about interactions. A modified handheld gripper records wrist RGB and samples two-channel TENG signals at 500 Hz. These recordings support VTLA-QA, comprising 100,000 QA instances after paraphrase expansion. Our framework combines visual QA supervised fine-tuning (SFT) with RGB-TENG alignment through shared EO-1 adapters, then applies Group Relative Policy Optimization (GRPO) to verifiable visual answers. The retained model learns robot actions from separate Mobile ALOHA demonstrations. On the held-out QA partition, SFT with Paired TENG improves multiple-choice accuracy from 60.10% for No TENG to 69.53%. Subsequent GRPO reaches 73.91%. Across four robot sorting settings, Paired TENG achieves a macro-averaged success rate (SR) of 66.25%, compared with 55.00% for No TENG. Deployment uses wrist RGB, instructions, and robot state. These results support a benefit from recorded TENG supervision under the evaluated training conditions, while leaving broader task generalization and the role of specific visual contact cues unresolved.**

I. Introduction

Robot manipulation requires interpreting how objects respond to contact. Distinguishing real from artificial fruit, for example, may involve observing deformation during probing. Vision-language-action (VLA) models connect visual and linguistic knowledge to robot control [1]–[4], while physical Question Answering (QA) provides supervision about objects and interactions [5], [6]. We investigate whether tactile recordings add useful supervision when deployment receives no tactile input.

A triboelectric nanogenerator (TENG) converts mechanical interaction into electrical signals [7]. Its time-varying responses provide a potential source of supervision beyond categorical QA targets. Prior studies use tactile experience to improve policies operating without tactile sensors [8]–[10]. Our question is whether raw-TENG alignment improves a vision-language model (VLM) already trained on the same QA targets, and whether the resulting representation transfers from handheld recordings to robot control.

We construct VTLA-QA from wrist RGB-TENG recordings acquired with a modified Pika gripper. Object records, event annotations, and measurements support questions with designated visual observations. Paraphrase expansion produces 100,000 QA instances, with all variants from each recording assigned to the same partition. Visual QA supervised fine-tuning (SFT) with Paired TENG updates shared EO-1 adapters through separate QA and alignment forward passes. After removing the signal branch, Group Relative Policy Optimization (GRPO) [11] adapts the visual answer policy using verifiable rewards. The retained model then learns actions from separate Mobile ALOHA demonstrations [12] (Fig. 1).

Under a common SFT recipe, Paired TENG improves multiple-choice accuracy by approximately 9.4 percentage points over No TENG. After GRPO and action fine-tuning, Paired TENG achieves a macro-averaged success rate (SR) of 66.25% across four sorting settings, compared with 55.00% for No TENG. We contribute a dataset connecting recorded RGB-TENG interactions to visually supported QA and a shared-adapter framework linking signal alignment, QA post-training, and action learning. Controlled comparisons of training routes and signal correspondence assess these components, with downstream evaluation on tactile-free robot sorting.

II. Related Work

Cross-modal learning aligns visual and other sensory representations [13], [14]. Visual-tactile learning uses robot and human interactions [15], [16], spans multiple tactile sensors [17]–[19], and incorporates language [20]–[22]. SuperTac incorporates triboelectric sensing into multimodal perception [23]. Tactile experience also supports robotic play [24], heterogeneous-sensor adaptation [25], and tactile-conditioned VLA policies [26], [27].

VITaL, HapticVLA, and TacImag address deployment without tactile sensors through visuo-tactile pretraining, distillation, and tactile prediction, respectively [8]–[10]. ACT and Diffusion Policy learn action sequences from demonstrations [28], [29]. We examine whether raw-TENG alignment adds value to the same visual QA supervision and transfers through shared adapters to tactile-free robot control.

III. Hardware System

The acquisition system combines wrist imaging, fingertip sensing, and gripper motion measurements to record handheld object interactions. Fig. 2 shows the modified assembly and fingertip sensor arrangement.

A. Instrumented handheld gripper

We modify a Pika handheld gripper [30] by integrating TENG and force-sensing resistor (FSR) elements into both fingertips. The TENG elements face the object-contact interface, with FSR elements mounted behind them. This arrangement records electrical responses and auxiliary load indications at the gripping surfaces while retaining the original wrist imaging and gripper motion measurements.

A wrist-mounted D405 camera supplies RGB observations of the gripper and manipulated object. HTC Vive tracking measures gripper position and orientation through a base-station system [31], providing six-degree-of-freedom handheld pose. A separate gripper-angle measurement records opening and closing. Together with calibrated gripper geometry, it supports fingertip-separation estimates for width questions. Pose and opening measurements provide annotation context for interpreting object and gripper motion.

B. Signal acquisition and temporal association

Two TENG channels are sampled at 500 Hz and retain their separate identities and signal polarities for representation alignment. FSR signals are also acquired at 500 Hz. Annotators inspect their load-related changes alongside RGB and TENG traces when interpreting contact events. FSR amplitudes are not treated as calibrated forces. These auxiliary measurements support annotation, while visual QA receives RGB and text.

Wrist RGB is acquired at approximately 30 frames per second with variable frame intervals. Recorded acquisition timestamps associate images with the 500-Hz signal streams. Window extraction selects images and signal samples over the corresponding recorded interval and preserves sample order within signal batches. Temporal pairing therefore uses the observed frame times without assuming uniformly spaced video frames. The resulting recordings provide the visual observations, paired signal windows, and auxiliary measurements used in dataset construction.

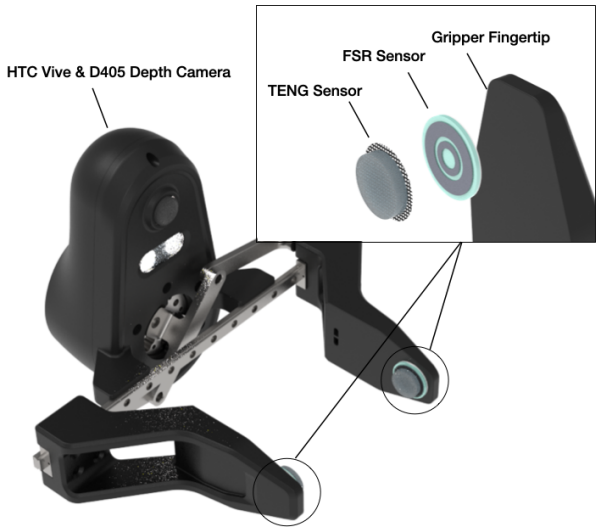

Fig. 2. Modified Pika gripper with fingertip TENG and FSR elements. Wrist imaging, Vive tracking, and gripper-angle measurements accompany the electrical recordings. The inset shows the fingertip assembly.

IV. Dataset Construction

A. Collection protocol

Operators probe objects by pressing, holding, lifting, rotating, and performing other interactions appropriate to their shape and material. An episode typically includes approach, contact onset, loading or holding, and release. Slip or shaking is annotated when observed. Annotators inspect RGB, TENG, and FSR traces to identify contact transitions and temporal relations, with gripper pose and opening measurements providing additional context. Fig. 3 illustrates an episode and representative questions. Robot demonstrations are collected separately for downstream sorting.

B. Item selection

Objects span materials, compliance, surface textures, geometry, and appearance, including cups, balls, sponges, textiles, natural materials, and real and artificial fruit.

Operator-defined soft/hard conditions describe the probing procedure, not calibrated material properties or force levels. Different orientations of one object retain the same physical identity. Downstream sorting settings and rules are defined in Sec. VI-D.

C. VTLA-QA design

VTLA-QA connects physical descriptions, interaction understanding, and instruction-conditioned decisions through twelve question families (Table I). Object records provide categorical attributes, event annotations establish interaction states and temporal targets, and calibrated gripper measurements support width estimates. Each source question

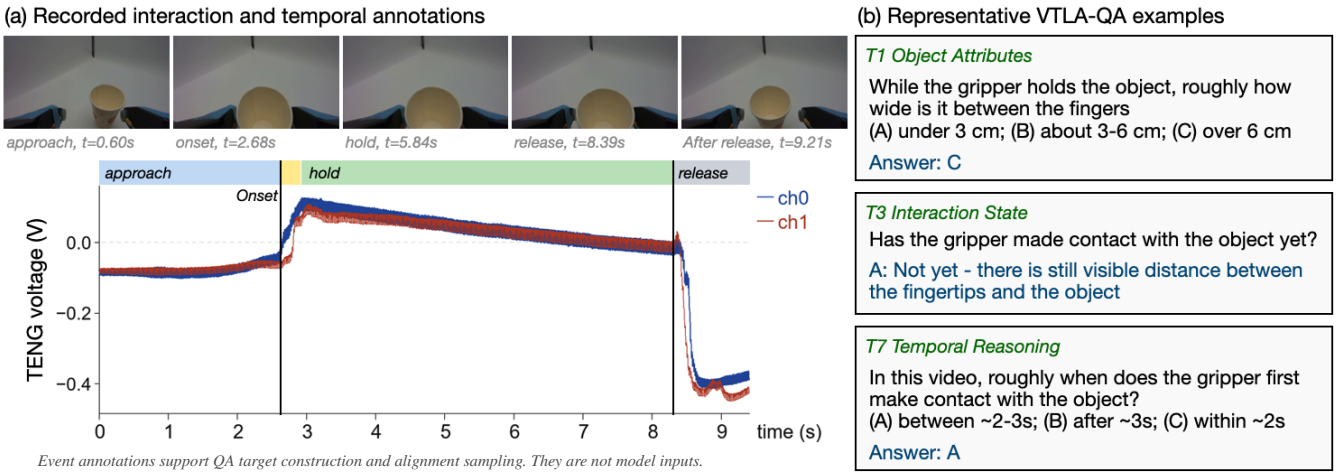

Fig. 3. Recorded interaction and representative VTLA-QA items. **(a)** Wrist RGB, two TENG channels, and temporal annotations from a cup interaction. **(b)** Questions about width, contact state, and event timing. Each question uses its designated observation. The contact-state example uses a pre-contact window.

TABLE I
VTLA-QA QUESTION FAMILIES.

ID	Question family	Scope
T1	Object attributes	Material, compliance, size
T2	Observability and uncertainty	Sufficiency of evidence
T3	Interaction understanding	Contact, deformation, motion
T4	Property comparison	Relative property judgments
T5	Outcome prediction	Possible consequences
T6	Action recommendation	Actions and adjustments
T7	Temporal reasoning	Event timing and order
T8	Visual disambiguation	Similar-looking objects
T9	Sorting decisions	Instruction-based destination
T10	Anomaly and risk assessment	Abnormal events and risks
T11	Counterfactual reasoning	Alternative conditions
T12	Description, analogy, and transfer	Summaries and related situations

is associated with a visual observation and reference answer. Explanations are restricted to the available evidence. Prediction and recommendation questions concern possible outcomes or actions. Sorting questions specify the class-to-box rule without revealing the object's class.

Approximately 20,000 source questions come from about 3,000 episodes involving 50 physical objects. We retain the original questions and add paraphrases that preserve their observations, meaning, answer formats, and reference answers. The expanded primary dataset contains 100,000 QA instances: 79,998 for training, 10,004 for validation, and 9,998 for test. Partitions are defined by source episode before expansion, with all variants remaining in the same partition. These counts represent linguistic instances rather than independently collected interactions.

Three answer formats are mutually exclusive. Multiple-choice items (MC) require one supplied option. Binary open-answer items (Bin) allow natural-language responses that a fixed, template-specific parser maps to one of two labels. Free-form open items (Open) comprise the remaining answers without deterministic correctness rules. All three formats supply SFT targets. Only MC and Bin enter GRPO and the primary accuracy metrics.

Each QA observation consists of eight RGB frames selected within its assigned temporal window using recorded timestamps. Questions assigned to contact, release, or whole-episode windows form the interaction subset, while pre-contact questions are evaluated separately.

Across all partitions, the dataset contains 50,662 MC, 29,975 Bin, and 19,363 Open instances. Fig. 4(a) shows their distribution across question families. Panels (b) and (c) summarize evidence timing relative to contact onset and question length by answer format. Overall, 45,376 instances use pre-contact observations and 54,624 use interaction observations.

V. VTLA ROBOT LEARNING

The route in Fig. 5 shares EO-1 adapters across QA SFT with Paired TENG, GRPO, and robot action learning.

A. Visual QA supervised fine-tuning

We use EO-1 3B [4] with its Qwen2.5-VL backbone [32]. Pretrained weights remain frozen during visual QA adaptation. Low-Rank Adaptation (LoRA) adapters [33] in the language backbone's attention and multilayer perceptron (MLP) projections are shared by QA and alignment and retained for subsequent stages.

Given RGB observations o, question q, and reference answer y, SFT minimizes answer-token cross-entropy $\mathscr{L}_{\mathrm{QA}}$ for the answer policy π_θ, where θ denotes adapter parameters. Prompt and padding tokens are excluded.

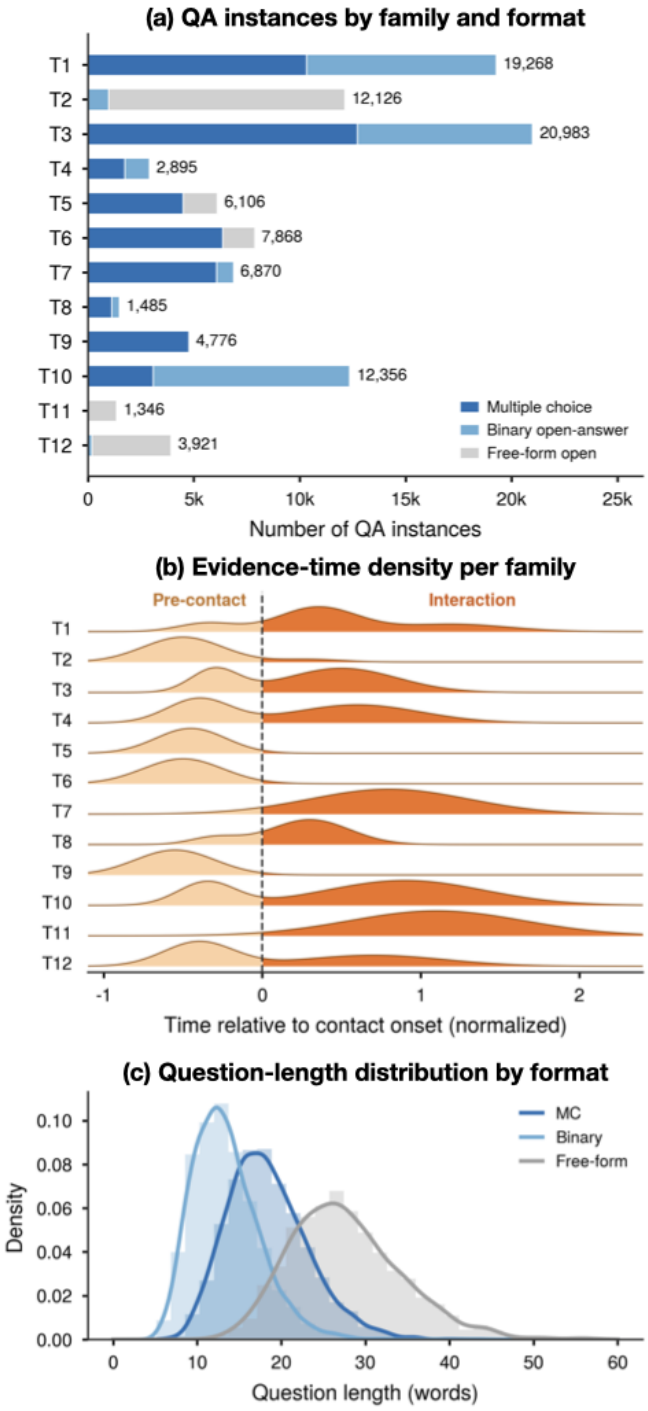

Fig. 4. Dataset statistics. **(a)** Distribution of 100,000 expanded QA instances across question families and answer formats. **(b)** Evidence-time density by family on a normalized time axis, with contact onset at zero. **(c)** Question-length distributions by answer format, measured in words. T1 to T12 follow Table I.

B. TENG Alignment

Each sample covers a two-second interval $[t_i - 2.0, t_i]$ from a training recording. We select eight RGB frames nearest to evenly spaced target timestamps and the corresponding two-channel TENG segment. Cutoff times are sampled uniformly where a complete interval is available. Alignment batches are drawn from paired recordings independently of QA paraphrase expansion.

EO-1 processes each RGB window with a fixed neutral prefix. Visual-token states are max-pooled and projected into a shared space. A one-dimensional TENG encoder uses channel widths 32/64/128, kernel sizes 7/5/3, stride 2, GELU activations, and channel-wise layer normalization. Temporal integration and projection produce a 128-dimensional embedding, matching the visual projection. Input normalization uses training data only, preserves channel identity and polarity, and masks missing samples.

Unit-normalized visual and signal embeddings v_i, s_j use similarity $v_i^{\mathsf{T}} s_j / \tau$, with $\tau = 0.07$. Symmetric InfoNCE [34] averages the two matching directions. Recorded RGB-TENG pairs provide positives. Eligible windows from other episodes and physical objects provide negatives. Joint training minimizes

$$\mathscr{L}_{\text{repr}} = \mathscr{L}_{\text{QA}} + \lambda \mathscr{L}_{\text{align}}, \qquad \lambda = 0.3. \tag{1}$$

QA and alignment use separate batches and forward passes through the shared adapters. Alignment also updates the TENG encoder and both projections. The visual representation used for alignment excludes answer targets. The signal encoder and alignment projections are discarded after SFT, while the updated EO-1 adapters are retained.

C. Visual QA reinforcement learning

GRPO [11] optimizes the EO-1 text-answer policy on training MC and Bin items. The GRPO-only route starts from Base EO-1. Each SFT+GRPO condition continues from its corresponding No TENG or Paired TENG SFT checkpoint.

For each question $x = (o, q)$, the policy samples $G = 8$ responses at temperature $T = 1.0$. A fixed, question-specific parser parse_q extracts an option or binary label. Given reference label $a^\star$, the reward is

$$r_i = \begin{cases} 1, & \text{parse}_q(y_i) = a^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The evaluator accesses $a^\star$. The policy sees only RGB and the question.

Rewards are normalized within each group as $\hat{A}_i = (r_i - \bar{r})/(\sigma_r + \varepsilon)$ and held constant during differentiation. Each freshly sampled batch receives one gradient update. With $\ell_{i,t}(\theta) = \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})$, the loss is

$$\mathscr{L}_{\text{GRPO}} = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left[-\hat{A}_i \ell_{i,t}(\theta) + \beta D_{i,t} \right], \tag{3}$$

where $D_{i,t} = \exp(\delta_{i,t}) - \delta_{i,t} - 1$, $\delta_{i,t} = \ell_{i,t}(\text{ref}) - \ell_{i,t}(\theta)$, and $\beta = 0.04$. Each route uses a frozen copy of its starting checkpoint as the reference policy. Groups with identical rewards have zero correctness advantage but retain the Kullback-Leibler (KL) penalty. Losses exclude prompt and padding tokens and are averaged over questions. Only the LoRA adapters are updated.

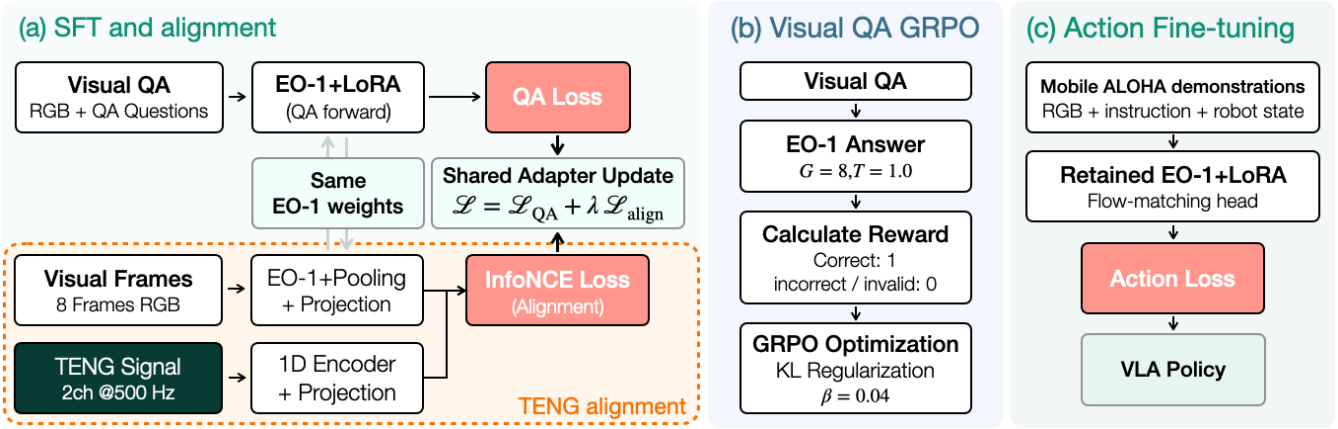

Fig. 5. Main training route. **(a)** QA SFT with Paired TENG updates shared EO-1 adapters through separate QA and alignment forward passes. **(b)** Visual QA GRPO optimizes answer generation using correctness rewards and KL regularization. **(c)** The retained model and a flow-matching action head learn from Mobile ALOHA demonstrations containing observations, instructions, robot states, and expert actions.

D. Action fine-tuning

The retained EO-1 model and adapters condition a flow-matching action head [3], [4]. Separate Mobile ALOHA demonstrations provide wrist RGB, instructions, robot states, and expert action chunks. Let c denote the input context and A an action chunk normalized using action-training statistics. We sample Gaussian noise ε and flow time $u \in [0,1]$, forming $A^u = (1-u)\varepsilon + uA$. The head $v_{\theta,\psi}$ learns the velocity $A - \varepsilon$ by minimizing

$$\mathcal{L}_{\text{act}} = \mathbb{E}\left[\frac{\left\|M \odot \left(v_{\theta,\psi}(A^u, u, c) - (A - \varepsilon)\right)\right\|_2^2}{\|M\|_1}\right], \qquad (4)$$

where M selects valid entries and $\odot$ denotes elementwise multiplication. The head and adapters are updated while pretrained backbone weights remain frozen. Numerical integration produces joint and gripper commands. The controller executes a chunk and replans from updated observations. TENG and FSR are not deployment inputs.

VI. Experiments and Discussion

We compare visual QA training routes, isolate the contribution of TENG alignment through SFT controls, and evaluate transfer to robot sorting.

A. Evaluation protocol

Table II reports held-out test results on identical questions across conditions. Trained QA conditions are reported for a single seed (29). The zero-shot baseline is evaluated once.

For MC and Bin, responses use unconstrained greedy decoding and are parsed into options or binary labels. Accuracy pools correct parsed-label counts over all evaluated questions, with invalid outputs counted as errors. We report overall, pre-contact, and interaction accuracy for each format. The pre-contact binary subset contains only interaction-understanding questions, so subgroup comparisons also reflect question composition. Scores are percentages, and differences are percentage points.

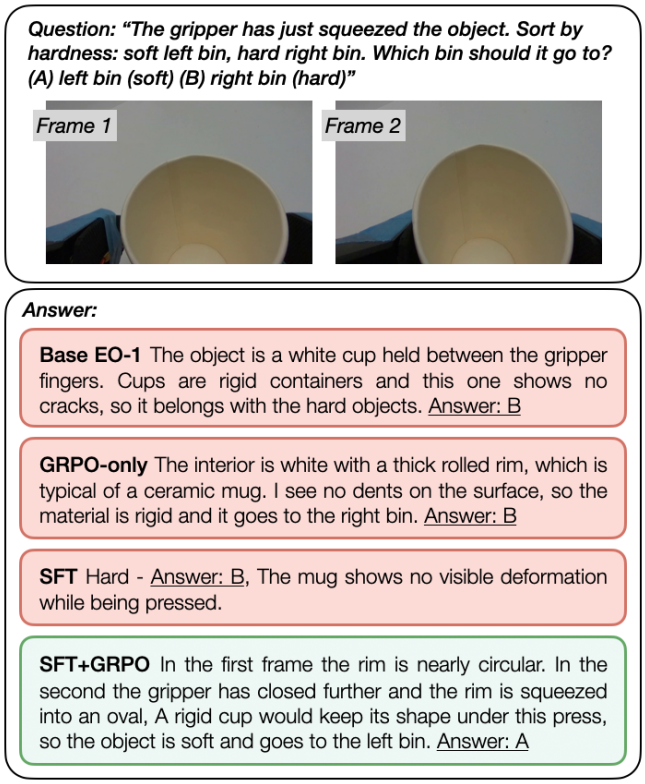

Fig. 6. Responses to a shared multiple-choice question. The question and responses are reproduced verbatim. Colors indicate option correctness. Explanations are shown for qualitative inspection. SFT and SFT+GRPO use Paired TENG, whereas Base and GRPO-only use No TENG.

During visual QA adaptation, LoRA uses rank 32, scaling factor 64, and dropout 0.05 in the attention projections (q, k, v, o) and MLP projections (gate, up, down), with approximately 59.87 million trainable parameters. Conditions using the same seed share adapter initialization. Each frame contributes 120 visual tokens.

SFT uses AdamW with effective batch size 16 (microbatch one), up to 2,000 updates, weight decay 0.01, and gradient-

TABLE II

TEST-SET VISUAL QA PERFORMANCE ACROSS TRAINING ROUTES AND TENG ALIGNMENT CONDITIONS. ACCURACIES ARE PERCENTAGES. TRAINED CONDITIONS USE SEED 29. ALIGNMENT IS APPLIED ONLY DURING SFT. EACH SFT+GRPO CONDITION CONTINUES FROM ITS CORRESPONDING SFT CHECKPOINT.

Training route	TENG Alignment	MC accuracy (%)			Binary open-answer accuracy (%)		
		Overall ↑	Pre-contact ↑	Interaction ↑	Overall ↑	Pre-contact ↑	Interaction ↑
Test questions (n)	N/A	4,999	2,686	2,313	2,974	209	2,765
Base (zero-shot)	No TENG	40.93	62.51	15.87	41.59	54.07	40.65
GRPO-only	No TENG	56.41	65.04	46.39	52.96	61.24	52.33
SFT	No TENG	60.10	66.50	52.70	55.50	69.90	54.40
SFT	Shuffled TENG	54.80	65.50	42.40	46.80	68.40	45.20
SFT	Paired TENG	69.53	69.21	69.91	72.56	73.68	72.48
SFT+GRPO	No TENG	64.70	69.30	59.30	57.60	76.07	56.20
SFT+GRPO	Paired TENG	**73.91**	**71.78**	**76.39**	**74.48**	**78.95**	**74.14**

norm clipping at 1.0. Learning rates are 10^{-4} for adapters and 3×10^{-4} for the TENG encoder and projections. A 60-update linear warm-up precedes cosine decay to 10% of the initial rate. Questions are sampled without replacement. After the first 500 updates, a fixed, stratified subset of 500 validation MC questions is evaluated every 250 updates. Training stops after three consecutive checks without an improvement of at least 0.5 percentage points. The best checkpoint is retained.

Each alignment update uses 16 pairs from eight objects and two episodes per object, with at least four eligible negatives per anchor. We fix $\lambda = 0.3$, selected from $\{0.03, 0.1, 0.3\}$ by validation MC accuracy in 1,000-update runs with seed 17. GRPO uses eight questions per update, a microbatch of four, learning rate 10^{-5}, gradient-norm clipping at 1.0, and up to 300 updates. Sampling uses top-$p = 1.0$ and response limits of 96 tokens for MC and 64 for Bin. Checkpoints are evaluated every 100 updates on the same validation subset, and the best is retained. Final model and hyperparameter selection use the full validation MC partition before test evaluation.

B. Visual QA post-training

In Table II, SFT+GRPO with Paired TENG achieves the highest observed accuracy in both answer formats, improving overall MC accuracy by 4.38 percentage points over SFT with Paired TENG. GRPO-only improves on Base but remains below SFT with Paired TENG. The additional MC improvement after GRPO is larger for interaction than pre-contact windows. These results compare complete training routes. The following SFT controls isolate alignment under a common adaptation procedure. Fig. 6 illustrates the original responses to one shared question.

C. Contribution of TENG supervision

The three SFT conditions in Table II share QA data, adapter initialization, visual preprocessing, optimizer settings, and update budgets. The No TENG condition uses QA supervision alone, Shuffled TENG aligns mismatched signal windows, and Paired TENG uses recorded RGB-TENG correspondence. Paired TENG and Shuffled TENG use identical alignment modules and settings.

For Shuffled TENG, each batch permutes complete signal windows across episodes without retaining an original pair. The permutation is resampled per batch without matching object identity or event type. Channel identity, temporal order, and validity masks are preserved. Reassigned windows serve as the contrastive positives. RGB observations and QA targets remain unchanged.

Paired TENG improves overall MC accuracy by approximately 9.4 percentage points over No TENG, with the larger difference in interaction windows. Binary accuracy follows the same ordering, and Shuffled TENG performs below No TENG.

D. Transfer to robot sorting

Robot experiments are conducted on a Trossen Mobile ALOHA platform [12]. Policies receive wrist RGB, instructions, and measured robot state. Four settings distinguish real/artificial fruit, original/imitation baseballs, glass/plastic cups, and 120-/180-grit sanding sponges. Artificial fruit, original baseballs, glass cups, and 120-grit sponges go to the gray box. The other class in each setting goes to the wooden box. Instructions specify this rule without revealing the current object's class.

We compare Base EO-1 with No TENG and Paired TENG initializations, both obtained by QA SFT followed by GRPO with matched QA data and update budgets. The adapted initializations correspond to the No TENG and Paired TENG SFT+GRPO rows in Table II. All three share the action-training recipe and episode partitions, with one policy per setting. Episode-disjoint training/validation splits contain 84/21 fruit, 75/18 baseball, 33/8 cup, and 48/12 sponge episodes. Validation supports checkpoint selection and offline diagnostics.

All action models use AdamW, batch size 8, and 16-step chunks. EO-1 uses two accumulation steps (effective batch size 16), learning rate 10^{-4}, weight decay 0.01, no gradient clipping, and seed 17. Its 30,000-update budget includes 1,500 warm-up updates. The learning rate halves after eight validation checks without improvement, at most three times. EO-1 has approximately 76.85 million trainable parameters.

TABLE III

TACTILE-FREE SORTING SR: SUCCESSES/TRIALS (%). ALL POLICIES UNDERGO ACTION TRAINING. SFT AND GRPO DENOTE VISUAL QA ADAPTATION. TENG ALIGNMENT IS APPLIED ONLY DURING SFT. MACRO SR AVERAGES THE FOUR SETTINGS.

Model	Fruit SR ↑	Baseball SR ↑	Cups SR ↑	Sponges SR ↑	Macro SR (%) ↑
$\pi_{0.5}$	10/20 (50)	10/20 (50)	9/20 (45)	8/20 (40)	46.25
OpenVLA-OFT	11/20 (55)	10/20 (50)	10/20 (50)	7/20 (35)	47.50
Base EO-1	9/20 (45)	9/20 (45)	8/20 (40)	6/20 (30)	40.00
EO-1 + SFT + GRPO (No TENG)	13/20 (65)	12/20 (60)	9/20 (45)	10/20 (50)	55.00
EO-1 + SFT + GRPO (Paired TENG, Ours)	**15/20 (75)**	**13/20 (65)**	**13/20 (65)**	**12/20 (60)**	**66.25**

Fig. 7. Selected sorting successes of the policy trained with Paired TENG and failures from comparison policies. Rows show fruit, baseball, cup, and sanding-sponge tasks. Main views are for visualization. Insets show wrist RGB. Labels describe individual trials. Table III reports aggregate success rates.

External configurations use $\pi_{0.5}$ [35] and OpenVLA-OFT [36] without dataset-specific QA post-training. Both use no gradient accumulation, gradient clipping at 10, and seed 1000. The $\pi_{0.5}$ action head is trained from scratch at constant learning rate 10^{-5} with weight decay 10^{-4}. OpenVLA-OFT freezes its visual encoder and trains its action expert with weight decay 10^{-10}, 1,000 warm-up updates, and cosine decay from 10^{-4} to 2.5×10^{-6} over 30,000 updates. Deployment inputs and evaluation rules are shared, while pretraining and adaptation differ.

Each policy receives 20 trials per setting, with 10 trials per class. A trial presents one object and two boxes. Success requires correct-box placement within 30 s. Fruit trials reuse five instances absent from both QA data and action demonstrations: three real bananas and two artificial replicas. Sponge trials alternate the two grit classes. Fig. 7 illustrates selected executions.

For setting k, $\mathrm{SR}_k = 100 S_k / N_k$, where S_k counts successes among N_k trials. Macro SR is the equally weighted mean

across the four settings. Table III reports success counts and rates, with Paired TENG achieving 66.25% Macro SR, 11.25 percentage points above No TENG. Class-level effects are examined in Sec. VI-E.

E. Discussion

Under matched SFT conditions, Paired TENG benefits interaction questions more than pre-contact questions, while Shuffled TENG reduces accuracy. This pattern supports the value of recorded correspondence, although it does not isolate which visual cues the model uses. GRPO yields further gains, especially on interaction questions.

After GRPO and action learning, Paired TENG achieves higher observed sorting success under the shared EO-1 recipe, but gains vary across classes. Improvements on artificial fruit and imitation baseballs accompany declines on real bananas and original baseballs. External VLA results provide context because architecture, pretraining, and adaptation differ.

VII. Limitations and Future Work

QA results use one seed, and each of four sorting settings has 20 trials, limiting precision and generalization. Fruit trials reuse five held-out instances (three real bananas and two artificial replicas), and paraphrases add no physical observations. Success scores conflate destination selection and execution.

Future work will expand objects, tasks, trials, and training seeds, separate decision and execution errors, and test contact sensitivity under controlled appearance. Further comparisons should match SFT and GRPO update budgets and extend the alignment-weight search beyond the tested range.

VIII. Conclusion

We presented shared VLM training with visual QA and recorded TENG alignment for tactile-free robot sorting. Controlled SFT comparisons support a benefit from signal correspondence. After GRPO and action learning, Paired TENG achieves 66.25% Macro SR versus 55.00% for No TENG. Establishing robustness across seeds and broader tasks remains open.

References

[1] A. Brohan, *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Proc. Conf. Robot Learn. (CoRL)*, 2023.

[2] M. J. Kim, *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[3] K. Black, *et al.*, "π_0: A Vision-Language-Action Flow Model for General Robot Control," in *Proc. Robot.: Sci. Syst. (RSS)*, 2025.

[4] D. Qu, *et al.*, "EO-1: An Open Unified Embodied Foundation Model for General Robot Control," arXiv:2508.21112, 2025.

[5] J. Gao, *et al.*, "Physically Grounded Vision-Language Models for Robotic Manipulation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.

[6] P. Sermanet, *et al.*, "RoboVQA: Multimodal Long-Horizon Reasoning for Robotics," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.

[7] F.-R. Fan, Z.-Q. Tian, and Z. L. Wang, "Flexible triboelectric generator," *Nano Energy*, vol. 1, no. 2, pp. 328–334, 2012.

[8] A. George, S. Gano, P. Katragadda, and A. B. Farimani, "VITaL Pretraining: Visuo-Tactile Pretraining for Tactile and Non-Tactile Manipulation Policies," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025, pp. 258–264.

[9] K. Gubernatorov, *et al.*, "HapticVLA: Contact-Rich Manipulation via Vision-Language-Action Model without Inference-Time Tactile Sensing," arXiv:2603.15257, 2026.

[10] Z. Zhang, *et al.*, "Imagining the Sense of Touch: Touch-Informed Manipulation via Imagined Tactile Representations," arXiv:2607.01684, 2026.

[11] Z. Shao, *et al.*, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," arXiv:2402.03300, 2024.

[12] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[13] A. Radford, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[14] R. Girdhar, *et al.*, "ImageBind: One Embedding Space To Bind Them All," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.

[15] J. Kerr, *et al.*, "Self-Supervised Visuo-Tactile Pretraining to Locate and Follow Garment Features," in *Proc. Robot.: Sci. Syst. (RSS)*, 2023.

[16] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, "Touch and Go: Learning from Human-Collected Vision and Touch," in *Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks*, 2022.

[17] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson, "Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[18] C. Higuera, *et al.*, "Sparsh: Self-Supervised Touch Representations for Vision-Based Tactile Sensing," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[19] C. Higuera, *et al.*, "Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation," arXiv:2506.14754, 2025.

[20] L. Fu, *et al.*, "A Touch, Vision, and Language Dataset for Multimodal Alignment," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[21] S. Yu, K. Lin, A. Xiao, J. Duan, and H. Soh, "Octopi: Object Property Reasoning with Large Tactile-Language Models," in *Proc. Robot.: Sci. Syst. (RSS)*, 2024.

[22] F. Yang, *et al.*, "Binding Touch to Everything: Learning Unified Multimodal Tactile Representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[23] S. Li, *et al.*, "Biomimetic multimodal tactile sensing enables human-like robotic perception," *Nature Sensors*, vol. 1, no. 1, pp. 52–62, 2026.

[24] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from Touch: Self-Supervised Pre-Training of Tactile Representations with Robotic Play," in *Proc. Conf. Robot Learn. (CoRL)*, 2023.

[25] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond Sight: Finetuning Generalist Robot Policies with Heterogeneous Sensors via Language Grounding," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025.

[26] Z. Cheng, *et al.*, "OmniVTLA: Vision-Tactile-Language-Action Models with Semantic-Aligned Tactile Sensing," arXiv:2508.08706, 2025.

[27] C. Morissette, *et al.*, "Tactile Modality Fusion for Vision-Language-Action Models," arXiv:2603.14604, 2026.

[28] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Proc. Robot.: Sci. Syst. (RSS)*, 2023.

[29] C. Chi, *et al.*, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in *Proc. Robot.: Sci. Syst. (RSS)*, 2023.

[30] AgileX Robotics, "Pika SDK," https://github.com/agilexrobotics/pika_sdk, accessed: Sep. 12, 2026.

[31] HTC Corporation, *VIVE Tracker (3.0) Developer Guidelines*, 1st ed., HTC Corporation, 2021, public release: Jan. 18, 2021. Accessed: Sep. 12, 2026.

[32] S. Bai, *et al.*, "Qwen2.5-VL Technical Report," arXiv:2502.13923, 2025.

[33] E. J. Hu, *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[34] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv:1807.03748, 2018.

[35] Physical Intelligence, *et al.*, "$\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization," arXiv:2504.16054, 2025.

[36] M. J. Kim, C. Finn, and P. Liang, "Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success," in *Proc. Robot.: Sci. Syst. (RSS)*, 2025.