SEEING HEAT, ACTING BY TOUCH: THERMALLY GROUNDED PROMPTING FOR VISION-TACTILE-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Tactile-Language-Action (VTLA) models augment Vision-Language-Action (VLA) policies with tactile feedback, enabling robots to respond to contact deformation, slip, and grasp stability during physical interaction. However, tactile sensing is available only after contact and cannot safely identify pre-contact target properties such as object temperature or electrical energization. A direct solution is to inject both thermal and tactile tokens into a pretrained action backbone. With limited robot demonstrations, however, jointly learning heterogeneous representations, cross-modal alignment, and continuous control from action supervision is data-inefficient and prone to RGB-dominated modality imbalance, potentially suppressing thermal and contact-dependent tactile cues. Scaling synchronized RGB–thermal–tactile–language–action data to alleviate this problem is also substantially more resource-intensive than collecting conventional RGB-based robot demonstrations. To address these challenges, we propose Thermo-VTLA, a hierarchical framework that separates thermal physical-state grounding from contact-rich VTLA control. A thermally grounded vision-language model receives paired RGB–thermal observations and the original instruction, and produces a constrained executable prompt. A prompt-conditioned VTLA controller then executes the command using RGB observations, robot state, and visuotactile feedback, without receiving thermal tokens. To effectively incorporate tactile feedback without disrupting the pretrained vision-language representation, we introduce a two-stage contact-aware tactile integration strategy. Contact-aware gating first determines tactile validity from the interaction state, after which gated fusion adaptively regulates the contribution of valid tactile representations to the action context. This decomposition preserves tactile sensing in the high-frequency control loop while moving thermal reasoning to an explicit semantic interface. We evaluate Thermo-VTLA on beverage retrieval, bread manipulation, and electrical-plug disconnection.

1 INTRODUCTION

Vision-Language-Action (VLA) models have emerged as a powerful method for embodied intelligence, enabling robots to execute complex manipulation tasks directly from visual observations and natural language instructions (Black et al., 2025b; Kim et al., 2024; Li et al., 2024b;a). By leveraging large-scale vision-language pretraining and flow/diffusion-based action heads, recent VLA models can map high-level instructions and RGB observations to low-level robot actions, demonstrating impressive generalization across diverse manipulation tasks (Black et al., 2025a; Shukor et al., 2025; Zitkovich et al., 2023; Chen et al., 2025).

Despite these advances, existing VLA models predominantly rely on RGB observations, which fundamentally limits their ability to achieve physically grounded interaction with the real world. While visual perception provides rich semantic information, it cannot directly reveal many intrinsic physical properties that are critical for manipulation, such as contact state, material compliance, texture, force, and temperature (Yue et al., 2026; Qian et al., 2026). Consequently, robots equipped with RGB-only VLA models may understand what an object is, but still lack sufficient understanding of how the object physically behaves during interaction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To enhance the physical understanding of VLA models, recent studies have explored Vision-Tactile-Language-Action (VTLA) frameworks by incorporating tactile sensing into VLA policies (Zhang et al., 2026a; Huang et al., 2026a; Cheng et al., 2026; Ye et al., 2025; Yuan et al., 2026). Tactile feedback provides direct physical interaction signals, including contact states, grasp stability, slip conditions, and object deformation, enabling more reliable closed-loop manipulation (Bi et al., 2026; Zhang et al., 2026b; Team et al., 2026). However, Tactile sensing does not remove every physical ambiguity. It becomes informative only after the gripper touches an object, whereas target selection often has to be completed before contact (Hao et al., 2026; Huang et al., 2026b). Two visually identical beverage cans can have different temperatures; similar bread samples can differ in whether they were recently heated; and an electrical component can be energized without an obvious visible change. An RGB-tactile controller can improve the grasp after contact, but it cannot safely determine which object satisfies a temperature-dependent instruction before approaching it.

Thermal infrared sensing supplies the missing pre-contact evidence by observing surface temperature and heat distribution without touching the scene. A direct approach is therefore to add both thermal and tactile encoders to a pretrained VLA and fuse their tokens with RGB, language, state, and action representations (Liu et al., 2026; Yu et al., 2026). This creates a demanding learning problem under the data scale available in robot manipulation. The policy must learn sensor-specific representations, align modalities with different spatial support and temporal validity, ground physical words such as *cold* or *powered*, and learn continuous actions from the same limited demonstrations. Action loss provides only indirect supervision for these alignment problems. Moreover, collecting large synchronized datasets requires calibrated RGB and thermal cameras, a gripper-mounted tactile sensor, robot trajectories, contact interactions, and repeated control of the objects' thermal state. Such data are costly to collect and difficult to scale.

The sensing roles suggest a factorization that reduces this burden. Thermal observations are most useful for deciding what object to manipulate, or whether the robot should act, before contact. Tactile observations are most useful for determining how to maintain and adjust the interaction after contact. Once the thermal evidence has been converted into an explicit spatial or behavioral instruction, the low-level controller can retain its RGB and tactile inputs without learning thermal-action alignment from every robot trajectory.

We propose Thermo-VTLA, a hierarchical framework that couples a thermally grounded reasoner with a prompt-conditioned VTLA controller. At the beginning of each episode, a VLM receives the external RGB image, the corresponding thermal image, and the user's instruction, and selects an executable prompt from a task-conditioned candidate set rather than generating an unrestricted plan. For example, *pick up the iced Coke* is converted into *pick up the right Coke*. The low-level controller then conditions on this prompt, external and wrist RGB observations, robot state, and visuotactile feedback to generate action chunks. Importantly, thermal information is removed only from the action backbone, whereas tactile sensing remains within the closed-loop controller to support contact-rich execution. A frozen tactile tokenizer encodes the visuotactile observation, contact-aware gating determines whether the tactile signal is physically valid, and gated fusion adaptively regulates its contribution to the vision-language action representation. This separation allows paired RGB–thermal scenes to support physical-state grounding and RGB–tactile robot demonstrations to train low-level control.

Our main contributions are summarized as follows:

- We introduce Thermo-VTLA, a hierarchical framework that uses an executable language prompt to separate thermal physical-state grounding from vision-tactile-language-action control.
- We propose a contact-aware tactile integration strategy that combines Contact-aware Gating and Gated Fusion to selectively activate tactile feedback and dynamically balance tactile information with pretrained VLM representations during manipulation.
- Real-world experiments across three physical-aware manipulation tasks demonstrate that Thermo-VTLA significantly improves task success rates over evaluated VLA baselines, highlighting the importance of thermal and tactile sensing for embodied physical intelligence.

2 Related Work

2.1 Vision-Language-Action Models

Recent advances in large-scale vision-language models have significantly accelerated the development of embodied intelligence, enabling robots to interpret natural language instructions and perform diverse manipulation tasks in open-world environments. Early VLA models, such as RT-2 (Zitkovich et al., 2023) and RT-X (O'Neill et al., 2024), demonstrated the feasibility of transferring knowledge from vision-language pretraining to robotic control by tokenizing robot actions and jointly modeling visual observations, language commands, and motor outputs. More recent approaches, including OpenVLA (Kim et al., 2024), RoboFlamingo (Li et al., 2024b), and π_0 (Black et al., 2025b), further improve generalization through large-scale robot demonstrations, stronger vision-language backbones, and diffusion/flow-based action generation. Despite their impressive capabilities, existing VLA models predominantly rely on RGB observations (Li et al., 2024a; Wen et al., 2025a;b), which limits their ability to perceive intrinsic physical properties that are not directly observable from appearance, such as temperature, contact state, material compliance, and surface texture. This limitation has motivated recent efforts toward physically grounded multimodal robotic foundation models that incorporate additional sensing modalities beyond vision.

2.2 Tactile Sensing for Robotic

Recent advances in tactile sensors and large-scale tactile datasets have shifted tactile perception from task-specific feature extraction toward generalizable multimodal representation learning (Luo et al., 2025). For example, Touch100K (Cheng et al., 2025) introduced a large-scale visuo-tactile dataset to facilitate learning of semantic and physical representations from tactile observations, while Octopi (Yu et al., 2024) demonstrated the potential of integrating visual and tactile modalities with language supervision through a unified Vision-Tactile-Language (VTL) (Fu et al., 2024) model. Building upon these efforts, subsequent works such as Octopi-1.5 (Yu et al., 2025) further explored scalable visual-tactile-language models, showing that tactile signals can be aligned with high-level semantic concepts and improve robotic understanding of physical interactions. Recently, these advances have inspired the integration of tactile representations into VLA frameworks, leading to emerging VTLA models that leverage tactile feedback for contact-aware manipulation (Zhang et al., 2026a; Huang et al., 2026a; Cheng et al., 2026; Ye et al., 2025; Yuan et al., 2026; Bi et al., 2026; Zhang et al., 2026b; Team et al., 2026). However, tactile sensing fundamentally depends on physical contact and therefore cannot provide information about non-contact physical properties, such as object temperature, before interaction. This inherent limitation motivates the incorporation of complementary sensing modalities to achieve more comprehensive physical understanding for embodied intelligence.

2.3 Thermal Sensing in Robotic Perception and Manipulation

Thermal infrared sensing provides complementary physical information that cannot be obtained from conventional RGB cameras, particularly for temperature-related perception and safety-critical robotic applications. Early robotic perception studies demonstrated the effectiveness of thermal imaging for human detection, object localization, and scene understanding under challenging illumination conditions (Song et al., 2023). More recent works have explored thermal sensing for robotic manipulation, including temperature-aware grasping, hot-object detection, and multimodal RGB–thermal perception, showing that thermal information can improve physical-state estimation beyond visual appearance (Cruz Ulloa et al., 2021; Ashqar et al., 2024; Im et al., 2025; Liu et al., 2026). Unlike tactile sensing, thermal perception enables non-contact measurement of object temperature and heat distribution before physical interaction, making it valuable for safe manipulation in real-world environments. However, existing thermal-based robotic systems mainly focus on standalone perception modules or task-specific control policies and rarely investigate how thermal evidence can support language-conditioned robot manipulation. A straightforward approach is to directly inject thermal tokens into a pretrained VLA backbone(Liu et al., 2026; Yu et al., 2026), but this requires the policy to jointly learn thermal representations, cross-modal alignment, and continuous action generation from limited robot demonstrations. In contrast, our framework semantically maps physical-state information encoded in thermal observations into executable prompts compatible with the language interface of pretrained VLA models. The proposed Thermo-VTLA controller then ex-

ecutes these prompts using vision and tactile feedback, enabling physical-property-aware robotic manipulation without directly injecting thermal tokens into the action policy.

3 METHOD

3.1 PROBLEM FORMULATION

Let l be an instruction containing a physical predicate, such as *pick up the iced Coke*. At the beginning of an episode, the high-level reasoner G_ϕ observes an external RGB image I_0^e, which provides a global view of the workspace, together with a synchronized thermal image I_0^{th}. It produces an executable prompt q from a finite candidate set $\mathcal{C}(l, \mathcal{S}_0)$ defined by the task and the current scene $\mathcal{S}_0$:

$$q \sim G_\phi \left(I_0^e, I_0^{th}, l; \mathcal{C}(l, \mathcal{S}_0) \right), \qquad q \in \mathcal{C}. \tag{1}$$

The candidate prompts refer to targets through properties observable by the low-level controller, such as *left*, *middle*, or *right*. The thermal image is used only by G_ϕ. To adapt the high-level reasoner to different task scenarios, only the contents of the candidate prompt set $\mathcal{C}$ need to be replaced.

At control time t, the VTLA policy receives the executable prompt, external and wrist RGB images, a visuotactile observation, and robot state:

$$O_t^{ctrl} = \{ I_t^e, I_t^w, T_t, s_t, q \}. \tag{2}$$

It predicts an action chunk $A_t = \{a_t, \ldots, a_{t+H-1}\}$:

$$A_t \sim \pi_\theta \left(I_t^e, I_t^w, T_t, s_t, q \right). \tag{3}$$

This factorization yields the hierarchical policy:

$$p(A_t \mid I_0^e, I_0^{th}, l, O_t^{ctrl}) = \sum_{q \in \mathcal{C}} p_\phi(q \mid I_0^e, I_0^{th}, l)\, p_\theta(A_t \mid O_t^{ctrl}, q). \tag{4}$$

Here, p_ϕ and p_θ are the conditional distributions induced by the high-level reasoner G_ϕ and the low-level policy π_θ, respectively. It exposes a discrete physical-grounding decision between sensing and control, enabling the two failure sources to be evaluated separately.

3.2 THERMAL-GROUNDED EXECUTABLE PROMPTING

As shown in Figure 1, the high-level module uses a thermal-capable VLM, denoted G_ϕ, to perform cross-modal physical grounding. The RGB and thermal images are synchronized and presented with explicit modality identifiers. Thermal values are converted to a three-channel visualization.

The reasoner is prompted with three fields: (1) the original user instruction; (2) the paired RGB and thermal observations; and (3) the allowed executable prompts. For the beverage task, an example input-output pair is:

> **Instruction:** Pick up the cold Coke and place it on the coaster.
> **Allowed outputs:** {Pick up the left Coke and place it on the coaster; Pick up the right Coke and place it on the coaster; }.
> **Output:** Pick up the right Coke and place it on the coaster.

The reasoner is queried once at episode initialization because the physical predicate determines target selection before motion begins. By translating thermal observations into a discrete semantic prompt rather than directly conditioning the low-level controller on raw thermal tokens, this design mitigates the degradation in manipulation accuracy caused by ambient-temperature shifts between training and inference, as well as variations in object temperature, thereby improving generalization across different thermal conditions.

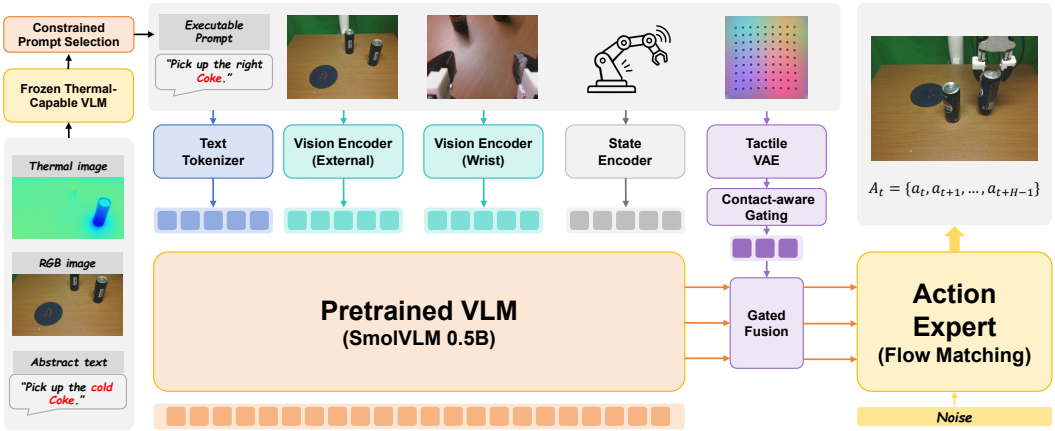

Figure 1: Overview of Thermo-VTLA, which uses thermal-grounded prompt selection for target identification and contact-aware visuotactile fusion for flow-matching action generation.

The VLM backbone in the low-level controller is pretrained primarily on RGB image–text data and lacks explicit thermal–language alignment. We therefore convey thermal evidence through executable language prompts rather than thermal feature vectors, allowing the controller to leverage its pretrained vision–language capabilities without learning this alignment from limited robot demonstrations. The prompts express the selected target through visually observable spatial relations, making the grounding decision directly inspectable and allowing a consistent command format to be used across different physical-state instructions. For example, the high-level reasoner converts *pick up the cold Coke* into *pick up the Coke on the right*, which the VTLA controller executes using RGB observations, robot proprioception, and tactile feedback during contact.

3.3 Prompt-Conditioned Vision-Tactile-Language-Action Controller

As shown in Figure 1, the VTLA controller is built on SmolVLA and consists of a pretrained SmolVLM backbone, a tactile pathway, and a flow-matching action expert. The refined prompt q, external RGB image I_t^e, wrist RGB image I_t^w, and state s_t are tokenized by their corresponding encoders and processed by the vision-language backbone:

$$H_t^{vl} = f_{\text{VLM}}\left(E_e(I_t^e), E_w(I_t^w), E_l(q), E_s(s_t)\right). \tag{5}$$

Here, f_{VLM} denotes the vision–language backbone, and H_t^{vl} denotes its output token representations. The tactile encoder, contact-aware gating, and gated fusion remain part of the policy. This architecture assigns thermal-language grounding to the high-level module and preserves touch where it directly supports motor execution. The action expert conditions on the vision-language representation and the fused tactile context. Because the executable prompts use the same linguistic form during demonstration collection and deployment, the controller is trained on exactly the interface it receives at test time. For each robot demonstration, the prompt label can be produced from the known target position and task template. The robot dataset therefore trains vision-tactile-language-action control without also having to learn thermal-language alignment.

3.4 Contact-aware Tactile Integration

A VQ-VAE encoder (Van Den Oord et al., 2017), pretrained on visuotactile images and subsequently frozen during VLA training, converts the visuotactile image T_t into a compact representation:

$$z_t^{tac} = E_{vq}(T_t), \qquad h_t^{tac} = P_{tac}\left(\text{Pool}(z_t^{tac})\right), \tag{6}$$

where P_{tac} is a trainable projection into the controller hidden dimension. We estimate contact from the displacement of N visual markers between consecutive tactile frames:

$$D_t = \frac{1}{N} \sum_{i=1}^{N} \left\| p_t^i - p_{t-1}^i \right\|_2, \tag{7}$$

and compute the binary contact state:

$$c_t = \mathbb{I}(D_t > \delta). \tag{8}$$

The valid tactile representation is $\tilde{h}_t^{tac} = c_t h_t^{tac}$. Let $h_t^{vl} = \mathrm{Pool}(H_t^{vl})$. A learned vector gate determines how much tactile information is injected:

$$g_t = \sigma\left(W_g[h_t^{vl}; \tilde{h}_t^{tac}] + b_g\right), \tag{9}$$

$$h_t^F = h_t^{vl} + g_t \odot \tilde{h}_t^{tac}. \tag{10}$$

Contact-aware activation answers *when* tactile evidence is meaningful, whereas the learned gate answers *how much* it should affect the current action. The fusion is applied after vision-language reasoning, avoiding interference with the pretrained semantic token sequence.

3.5 Flow-Matching Action Generation and Optimization

Following SmolVLA (Shukor et al., 2025), the action expert models a conditional velocity field for continuous action chunks. Given the demonstrated action A_t and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we sample $\tau \sim \mathcal{U}(0, 1)$ and construct

$$A_t^\tau = \tau A_t + (1 - \tau)\epsilon. \tag{11}$$

The target velocity under this interpolation is $A_t - \epsilon$. The training objective is

$$\mathcal{L}_{FM} = \mathbb{E}_{A_t, \epsilon, \tau}\left[\left\| v_\theta(A_t^\tau, \tau, H_t^{vl}, h_t^F) - (A_t - \epsilon) \right\|_2^2\right]. \tag{12}$$

At inference, the action sample is initialized from Gaussian noise and integrated along the learned vector field. The pretrained RGB encoders, language backbone, and tactile VQ-VAE are frozen. The tactile projection, fusion gate, state projection, and action expert are optimized:

$$\theta_{train} = \{P_{tac}, W_g, b_g, E_s, \pi_{expert}\}. \tag{13}$$

The high-level reasoner remains frozen across all tasks and requires neither training nor fine-tuning. Adapting the reasoner to a new task only requires updating the task instruction and the corresponding set of allowed executable prompts. Given paired RGB–thermal observations, the same VLM selects one prompt from this task-specific candidate set. Robot demonstrations are used solely to train the low-level VTLA controller to execute the resulting prompts.

4 Experiments

4.1 Experimental Setup

We evaluate Thermo-VTLA on three real-robot tasks: selecting a cold Coke from cold and room-temperature candidates (task1), selecting room-temperature bread from heated and room-temperature candidates (task2), and unplugging the powered phone charger from powered and unpowered candidates (task3). In each task, the two visually similar objects are randomly exchanged between fixed left and right positions. The frozen high-level VLM receives the initial RGB–thermal observations and the physical-state instruction, and selects either a left-target or right-target executable prompt.

For each task, we collect 50 trajectories with the target on the left and 50 with the target on the right, resulting in 100 demonstrations per task and 300 demonstrations in total. These demonstrations are used only to train the multi-task VTLA controller; the high-level VLM requires no training or fine-tuning. For each method, we conduct 20 independent real-robot trials on each task, including 10 trials with the target on the left and 10 trials with the target on the right. The same randomized target arrangements and initial conditions are used across all compared methods to ensure a fair comparison. We report Prompt Grounding Accuracy (PGA), and Task Success Rate (TSR). Detailed hardware configurations, candidate prompts, dataset contents, and training hyperparameters are provided in the Appendix.

4.2 Comparison with VLA Models

We compare Thermo-VTLA with $\pi_{0.5}$, SmolVLA, and a thermal-free variant of our framework. The first two baselines receive RGB observations, robot states, and the original physical-state instruction. Thermo-VTLA w/o Thermal retains the executable-prompt interface and the complete vision-tactile controller, but its high-level reasoner receives RGB without thermal observations.

Table 1: Comparison with VLA methods on three real-world physically grounded manipulation tasks. The best results are highlighted in bold.

Method	Metric	Beverage	Bread	Charger	Avg.
$\pi_{0.5}$	TSR	50.0	45.0	45.0	46.7
SmolVLA	TSR	40.0	40.0	30.0	36.7
Ours w/o Thermal	PGA	45.0	50.0	50.0	48.3
	TSR	40.0	40.0	30.0	36.7
Thermo-VTLA	**PGA**	**100.0**	**100.0**	**100.0**	**100.0**
	TSR	**100.0**	**85.0**	**95.0**	**93.3**

Results and analysis. As shown in Table 1, Thermo-VTLA achieves an average TSR of **93.3%**, outperforming $\pi_0.5$ and SmolVLA by **46.6** and **56.6** percentage points. Although the VLA baselines can learn the demonstrated actions, RGB observations do not reveal which candidate satisfies the requested physical state. Removing thermal input reduces PGA from **100.0%** to **48.3%** and average TSR by **56.6** percentage points. This result shows that tactile feedback supports execution after contact but cannot resolve target ambiguity beforehand. Thermo-VTLA combines thermal target grounding with vision-tactile execution to address both stages.

4.3 Comparison with Direct Multimodal Token Injection

We compare our hierarchical design with three direct-injection variants built on the same SmolVLA backbone. Thermal Tokens processes thermal images using an additional encoder with the same architecture and initialization as the RGB encoder. Tactile Tokens directly appends projected tactile representations to the VLM token sequence. Thermal+Tactile Tokens injects both modalities into the pretrained VLM. These variants receive the original physical-state instruction and directly predict actions.

Table 2: Comparison with direct multimodal token-injection strategies. The best results are highlighted in bold.

Method	Metric	Beverage	Bread	Charger	Avg.
Thermal Tokens	TSR	50.0	35.0	45.0	43.3
Tactile Tokens	TSR	45.0	45.0	30.0	40.0
Thermal + Tactile Tokens	TSR	45.0	45.0	45.0	45.0
Thermo-VTLA	**PGA**	**100.0**	**100.0**	**100.0**	**100.0**
	TSR	**100.0**	**85.0**	**95.0**	**93.3**

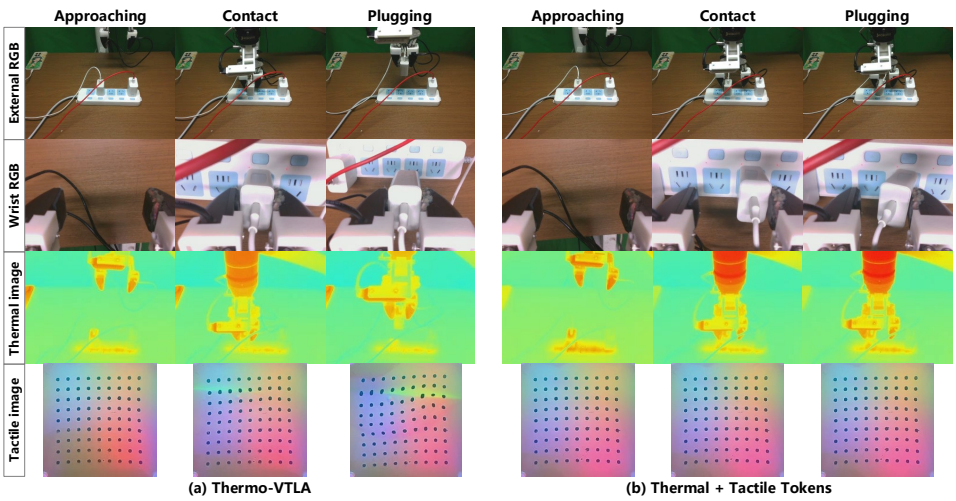

Figure 2: Qualitative comparison between direct RGB–thermal–tactile token injection and Thermo-VTLA in Task3. (See the supplementary video for further experimental details.)

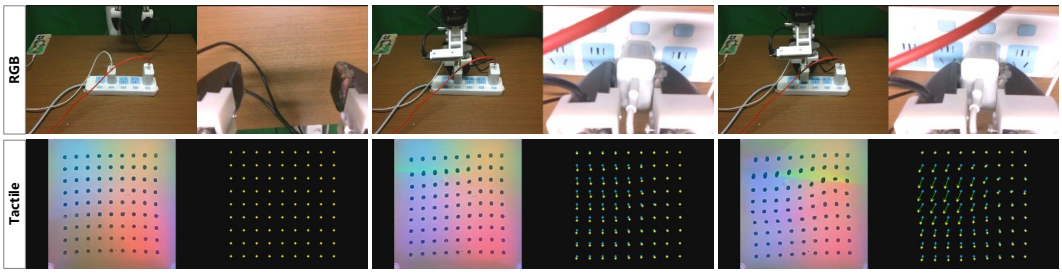

Figure 3: Evolution of visual and tactile observations during manipulation. The sequence shows the external RGB view, wrist RGB view, visuotactile image, and tracked-marker displacement field from pre-contact motion to stable physical contact.

Results and analysis. Table 2 shows that Thermo-VTLA outperforms the strongest direct-injection variant by **48.3** percentage points on average. Direct tactile tokens cannot resolve physical states before contact, while direct thermal tokens require an RGB-pretrained action model to learn thermal representations and thermal–language–action alignment from limited robot demonstrations. Injecting both modalities further increases the heterogeneous alignment burden and does not guarantee their effective utilization.

Our method instead maps physical-state information in thermal observations into executable language prompts compatible with the VLA interface. Tactile information is then incorporated only within the low-level controller, where it directly supports physical interaction. The results demonstrate the advantage of assigning thermal and tactile sensing to their appropriate stages rather than injecting all modality tokens into a single action backbone. Figure 2 provides a qualitative comparison between direct multimodal token injection and our hierarchical framework. Although the thermal observations contain a clear physical cue for identifying the powered charger, the direct-injection policy fails to consistently translate this cue into target-specific robot actions. This is because the policy must align heterogeneous thermal features and sparse, contact-dependent tactile features with RGB and language using only action supervision; moreover, tactile tokens are largely uninformative before contact and therefore cannot assist pre-contact target selection. In contrast, Thermo-VTLA first maps the thermal evidence to an explicit target-resolved prompt and introduces tactile information only when physical contact becomes valid, resulting in more accurate target-directed motion and more stable contact-rich execution. This comparison shows that the improvement arises from the role-aware decomposition of thermal grounding and tactile control, rather than merely from providing additional sensor tokens to the action backbone.

Table 3: Ablation of contact-aware tactile integration. The best results are highlighted in bold.

Variant	Metric	Beverage	Bread	Charger	Avg.
Ours w/o Tactile	TSR	80.0	70.0	70.0	73.3
Ours w/ CAG Only	TSR	90.0	70.0	80.0	80.0
Ours w/ GF Only	TSR	85.0	75.0	85.0	81.7
Ours All	**TSR**	**100.0**	**85.0**	**95.0**	**93.3**

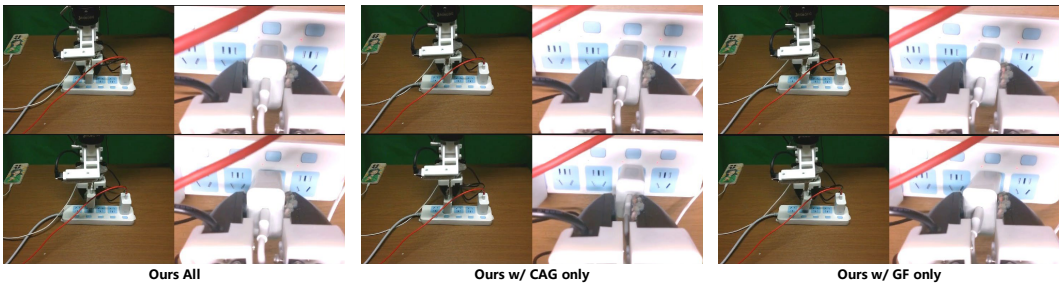

Figure 4: Qualitative ablation of Contact-aware Gating (CAG) and Gated Fusion (GF).

4.4 Ablation of Contact-Aware Tactile Integration

We evaluate four controller variants: removing tactile feedback, retaining only Contact-aware Gating (CAG), retaining only Gated Fusion (GF), and combining both modules. All variants receive the ground-truth executable prompt so that high-level grounding errors do not affect the tactile evaluation. We therefore report TSR in this experiment. Figure 3 visualizes the evolution of the visuotactile image and marker-displacement field during manipulation, showing that structured tactile deformation emerges only after physical contact is established.

Results and analysis. As shown in Table 3, removing tactile feedback decreases the average TSR from **93.3%** to **73.3%**, confirming that correct target grounding alone is insufficient for reliable physical execution. CAG and GF individually improve TSR by **6.7** and **8.4** percentage points, respectively.

CAG determines whether tactile observations correspond to valid physical contact, thereby suppressing non-contact variation. GF determines how strongly valid tactile information should influence the action representation. Combining both components achieves the highest TSR, demonstrating that contact-aware activation and adaptive fusion provide complementary benefits for contact-rich VTLA control. The qualitative comparison in Figure 4 further supports this conclusion. The complete model maintains more stable gripper–object alignment and contact throughout manipulation, whereas the CAG-only and GF-only variants exhibit less stable contact and larger execution deviations. These results demonstrate that combining CAG with GF improves grasp stability rather than merely increasing the amount of tactile information provided to the controller.

5 Conclusion

Thermo-VTLA converts RGB–thermal observations and an abstract physical instruction into a constrained executable prompt, then performs contact-aware VTLA control using RGB, robot state, and tactile feedback. The hierarchy preserves tactile sensing within the closed-loop controller while removing thermal-token alignment from robot policy learning. By assigning thermal perception to physical-state grounding and tactile perception to contact-dependent control, the proposed framework reduces the multimodal learning burden on limited robot demonstrations and provides a practical alternative to monolithic multimodal fusion.

AI USE STATEMENT

In this work, large language models (LLMs) were used to polish the manuscript and improve its readability, assist in retrieving relevant literature, and support code development and optimization. They were not involved in formulating the research ideas or contributions, designing the experiments, or analyzing and interpreting the experimental results. All substantive scientific content and claims were independently developed by the authors. The authors reviewed all LLM-assisted text for accuracy and originality and manually inspected and tested all LLM-assisted code. The authors take full responsibility for the final content of this work.

REFERENCES

Huthaifa I Ashqar, Taqwa I Alhadidi, Mohammed Elhenawy, and Nour O Khanfar. Leveraging multimodal large language models (mllms) for enhanced object detection and scene understanding in thermal images for autonomous driving systems. *Automation*, 5(4):508–526, 2024.

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Shou Zheng, and Harold Soh. Vla-touch: Enhancing vision-language-action model with dual-level tactile feedback. *IEEE Robotics and Automation Letters*, 11(7):8487–8494, 2026. doi: 10.1109/LRA.2026.3692345.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *RSS 2025*, 2025b.

Hao Chen, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Renrui Zhang, Xiaoqi Li, Xiao He, Yandong Guo, Chi-Wing Fu, Shanghang Zhang, and Pheng-Ann Heng. Fast-in-slow: A dual-system VLA model unifying fast manipulation within slow reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4asFznbzJg.

Ning Cheng, Jinan Xu, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *Information Fusion*, 124:103305, 2025.

Zhengxue Cheng, Yiqian Zhang, Anni Tang, Keyu Wang, Wenkang Zhang, Haoyu Li, Hengdi Zhang, and Li Song. Omnivtla: Vision-tactile-language-action models with semantic-aligned tactile sensing. *IEEE Robotics and Automation Letters*, 11(9):10146–10153, 2026. doi: 10.1109/LRA.2026.3710060.

Christyan Cruz Ulloa, Guillermo Prieto Sánchez, Antonio Barrientos, and Jaime Del Cerro. Autonomous thermal vision robotic system for victims recognition in search and rescue missions. *Sensors*, 21(21):7346, 2021.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.),

Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 14080–14101. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fu24b.html.

Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: tactile-language-action model for contact-rich manipulation. *Robot Learning*, 3(1), 2026. doi: 10.55092/rl20260001. URL https://doi.org/10.55092/rl20260001.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-VLA: Unlocking vision-language-action model's physical knowledge for tactile generalization, 2026a. URL https://openreview.net/forum?id=uhB3pbJpRm.

Yuzhe Huang, Pei Lin, Wanlin Li, Daohan Li, Jiajun Li, Jiaming Jiang, Chenxi Xiao, and Ziyuan Jiao. Taf-vla: Tactile-force alignment in vision-language-action models for force-aware manipulation. *CoRR*, abs/2601.20321, January 2026b. URL https://doi.org/10.48550/arXiv.2601.20321.

Seong-Min Im, Byeong-Sun Park, Jaehwan Jang, Sungeun Hong, Changjoo Nam, Young Tack Lee, and Min-gu Kim. Simultaneous in-hand shape and temperature recognition using flexible multilayered sensor arrays for sense-based robot manipulation. *Advanced Sensor Research*, 4(7): 70004, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ZMnD6QZAE6.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=lFYj0oibGR.

Xingyuming Liu, Ruichun Ma, Heyu Guo, Qixiu Li, Qingwen Yang, Lin Luo, Shiqi Jiang, Chenren Xu, Jiaolong Yang, and Baining Guo. Musevla: An adaptive multimodal sensing vision-language-action model for robotic manipulation. *arXiv preprint arXiv:2606.17598*, 2026.

Shan Luo, Nathan F. Lepora, Wenzhen Yuan, Kaspar Althoefer, Gordon Cheng, and Ravinder Dahiya. Tactile robotics: An outlook. *IEEE Transactions on Robotics*, 41:5564–5583, 2025. doi: 10.1109/TRO.2025.3608686.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra

Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi Jim Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick Tree Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Kun Qian, Zhuoyang Zhang, Ling Tong, and Shan Luo. Explorative grasping referred objects with weak stiffness via vision-language-tactile fusion. *IEEE Transactions on Industrial Electronics*, 73(7):10229–10241, 2026. doi: 10.1109/TIE.2026.3658641.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Kechen Song, Ying Zhao, Liming Huang, Yunhui Yan, and Qinggang Meng. Rgb-t image analysis technology and application: A survey. *Engineering Applications of Artificial Intelligence*, 120: 105919, 2023.

NeoteAI Team, Fudan TEAI Team, et al. n_0-vtla: Scaling vision-tactile-language-action model with latent tactile tokens. *arXiv preprint arXiv:2607.23782*, 2026.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. DexVLA: Vision-language model with plug-in diffusion expert for general robot control. In *9th Annual Conference on Robot Learning*, 2025a. URL https://openreview.net/forum?id=RFmezNsPWV.

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 10(4):3988–3995, 2025b.

Guo Ye, Zexi Zhang, Xu Zhao, Shang Wu, Haoran Lu, Shihan Lu, and Han Liu. Learning to feel the future: Dreamtacvla for contact-rich manipulation. *CoRR*, abs/2512.23864, December 2025. URL https://doi.org/10.48550/arXiv.2512.23864.

Dian Yu, Qingchuan Zhou, Bingkun Huang, Majid Khadiv, and Zewen Yang. Safe-night vla: Seeing the unseen via thermal-perceptive vision-language-action models for safety-critical manipulation. *arXiv preprint arXiv:2603.05754*, 2026.

Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. In *Robotics: Science and Systems*, 2024. URL https://doi.org/10.15607/RSS.2024.XX.066.

Samson Yu, Lin Kelvin, and Harold Soh. Demonstrating the Octopi-1.5 Visual-Tactile-Language Model. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.058.

Haoran Yuan, Weigang Yi, Zhenyu Zhang, Wendi Chen, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, et al. Vtam: Video-tactile-action models for complex physical interaction beyond vlas. *arXiv preprint arXiv:2603.23481*, 2026.

Zhaokun Yue, Ling Tong, and Kun Qian. Pocodp3: Pose- and contact-aware visual-tactile policy for contact-rich 3d manipulation. *IEEE Robotics and Automation Letters*, 11(3):2434–2441, 2026. doi: 10.1109/LRA.2026.3653314.

Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *Biomimetic Intelligence and Robotics*, pp. 100333, 2026a.

Xidong Zhang, Yichi Zhang, Jiaxin Shi, Fucai Zhu, Siyu Zhu, Michael Yu Wang, Xiaojun Wu, and Weihao Yuan. Unitacvla: Unified tactile understanding and prediction in vision language action models. *arXiv preprint arXiv:2606.31723*, 2026b.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R Sanketi, Grecia Salazar, Michael S Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J Joshi, Alex Irpan, brian ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=XMQgwiJ7KSX.

A Additional experimental settings

A.1 Robot Setup

All experiments are conducted on the real-robot platform shown in Fig. 5. The platform consists of a UR5 robotic arm equipped with an Inspire EG2-4C2 gripper and a ViTai visuotactile sensor. Visual observations are captured by an external Intel RealSense L515 camera and a wrist-mounted Intel RealSense D435 camera, while a fixed thermal infrared camera observes the temperature distribution of the workspace.

The external RGB camera provides a global view for scene understanding and target localization, whereas the wrist-mounted RGB camera captures local visual details during grasping and manipulation. The thermal infrared camera provides non-contact physical-state information that is not directly observable from RGB appearance, such as object temperature and heat generated by electrical activity. The ViTai sensor records contact deformation after the gripper interacts with an object, providing time-varying tactile feedback for contact-rich control.

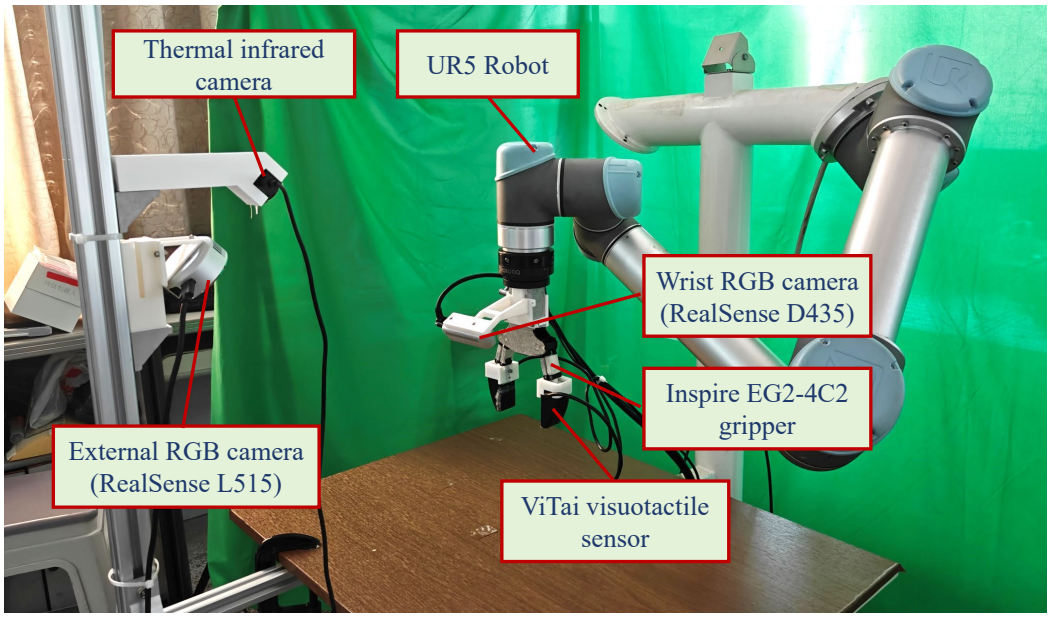

Figure 5: Real-robot platform used in the Thermo-VTLA experiments, including a UR5 robot, an Inspire EG2-4C2 gripper, external and wrist-mounted RGB cameras, a thermal infrared camera, and a ViTai visuotactile sensor.

At the beginning of each episode, the paired external RGB and thermal observations are provided to the frozen thermal-capable VLM to resolve the physical state specified by the original instruction and generate a constrained executable prompt. During robot execution, the VTLA controller receives the executable prompt together with synchronized external RGB, wrist RGB, robot proprioceptive state, and visuotactile observations at the control rate. Thermal observations are not provided to the low-level controller and are not included in VTLA policy training.

A.2 Training Details

The external Intel RealSense L515 camera and the wrist-mounted Intel RealSense D435 camera capture RGB-D observations at a resolution of 640×480 pixels. Although depth observations are available, only the RGB channels are provided to the VTLA controller. The thermal infrared camera produces images at its native resolution of 256×192 pixels, which are used only by the frozen high-level reasoner for physical-state grounding. The ViTai visuotactile sensor outputs tactile images with a resolution of 256×256 pixels and tracks a 9×9 array of two-dimensional marker displacements. The tactile images are encoded by a VQ-VAE pretrained on visuotactile observations and subsequently frozen during VLA training, whereas the marker displacements are used by Contact-aware Gating to determine whether physical contact has been established and, consequently, whether the tactile features should be activated.

Although visuotactile sensors are mounted on both sides of the gripper, the policy uses the tactile image and marker-displacement array from only one sensor during both training and inference. This setting reduces the tactile input dimensionality and inference cost while preserving the contact feedback required for the evaluated manipulation tasks. The same sensor configuration is used for all methods and ablation variants.

For each task, we collect 50 expert demonstration trajectories for each target position, resulting in 100 trajectories per task. Each controller training sample consists of an executable prompt, external and wrist RGB observations, robot proprioceptive state, the single-sensor tactile observation, and the corresponding demonstrated action sequence. The controller predicts an action chunk

$$A_t = \{a_t, a_{t+1}, \ldots, a_{t+H-1}\},$$

Table 4: Original physical-state instructions and candidate executable prompts for the three real-robot tasks.

Task	Original instruction	Candidate executable prompts $\mathcal{C}$
Beverage	*Grasp the cold Coke and place it on the coaster.*	**Left:** *Grasp the Coke on the left and place it on the coaster.* **Right:** *Grasp the Coke on the right and place it on the coaster.*
Bread	*Grasp the room-temperature bread and place it on the plate.*	**Left:** *Grasp the bread on the left and place it on the plate.* **Right:** *Grasp the bread on the right and place it on the plate.*
Charger	*Unplug the powered phone charger.*	**Left:** *Unplug the phone charger on the left.* **Right:** *Unplug the phone charger on the right.*

where the action horizon is set to $H = 50$.

The high-level thermal-capable VLM remains frozen and requires neither training nor task-specific fine-tuning. Task adaptation is performed only by changing the original instruction and its task-conditioned set of allowed executable prompts. For the low-level VTLA controller, the pretrained RGB encoders, vision-language backbone, and tactile VQ-VAE are frozen. For the direct token injection baselines, we use the same tactile encoder as in Thermo-VTLA. Specifically, tactile observations are encoded by the tactile VQ-VAE pretrained on visuotactile data, whose parameters remain frozen during VLA training. For thermal observations, we employ an encoder with the same architecture as the RGB visual encoder. Unlike the tactile VQ-VAE, the thermal encoder is not frozen and is trained jointly with the VLA model.

All VLA models are trained on an NVIDIA GeForce RTX 5090D GPU, each equipped with 32 GB of memory. All proposed variants and comparison methods are trained or fine-tuned on the same expert demonstration data for 80,000 optimization steps with a batch size of 64. The same action horizon, input observations, training split, and optimization budget are used across methods unless a modality is explicitly removed by the corresponding baseline or ablation.

B Task Instructions and Executable Prompts

Table 4 presents the original physical-state instructions and the corresponding candidate executable prompts used in the three tasks. In all experiments, we use `GPT-5.6-sol` as the high-level thermal-grounded VLM G_ϕ. The model is used in inference-only mode without any task-specific training or fine-tuning. At the beginning of each episode, G_ϕ identifies the target from the paired RGB–thermal observations and selects the appropriate left- or right-resolved prompt from the task-specific candidate set $\mathcal{C}$. The selected executable prompt is then provided to the low-level VTLA controller throughout the episode. For a new task scenario, only the contents of $\mathcal{C}$ need to be replaced; no additional training or fine-tuning of G_ϕ is required.

C Visualization of Multimodal Observations

Figures 6–8 visualize the multimodal observations before and after physical contact in the three manipulation tasks. All figures follow the same layout. From left to right, the columns show the thermal image, external RGB image, wrist RGB image, and visuotactile image. The first row presents the observations before grasping, whereas the second and third rows present the corresponding observations after the gripper has established contact with the target object.

Before grasping, the two candidate objects have similar RGB appearances, and the visuotactile sensor remains in its no-contact state. Consequently, the RGB and tactile observations cannot directly determine which object is referred to by a physical-state instruction such as *the cold Coke*, *the room-temperature bread*, or *the powered charger*. In contrast, the thermal observation reveals the physical-state difference between the candidates, allowing the high-level VLM to resolve the referent in the instruction and convert it into a left- or right-specific executable prompt.

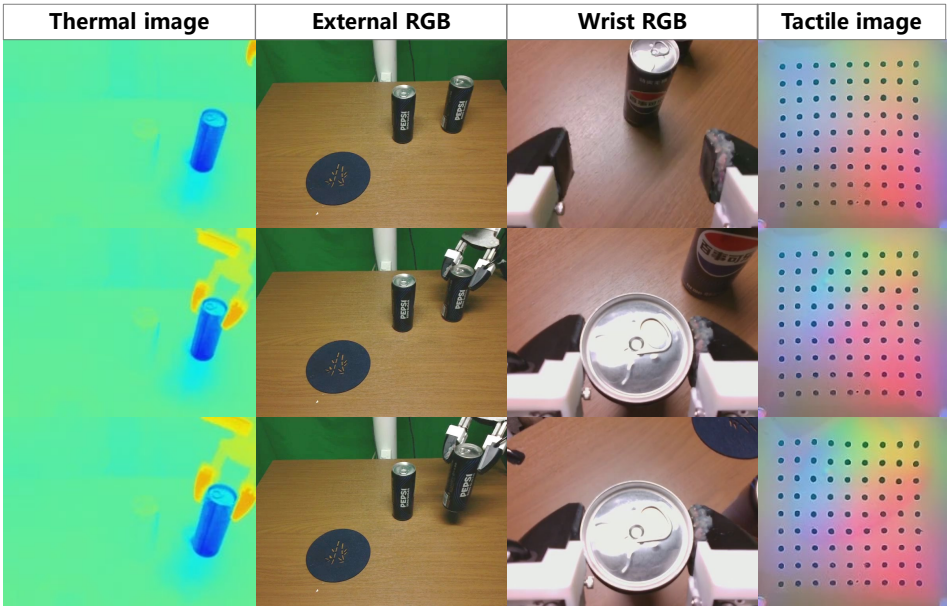

Figure 6: Multimodal observations in the cold-beverage retrieval task. Before grasping, the thermal image identifies the cold Coke referred to by the instruction. After grasping, contact deformation and tactile-marker displacement become clearly observable.

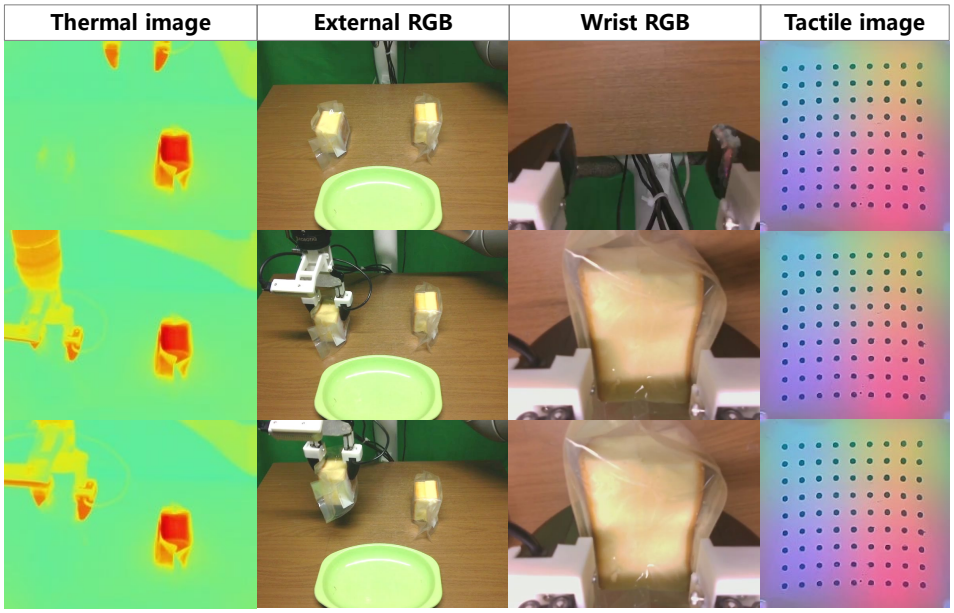

Figure 7: Multimodal observations in the bread-manipulation task. Before grasping, the thermal image identifies the room-temperature bread specified by the instruction. After grasping, the visuo-tactile image and marker displacement reflect the established physical contact.

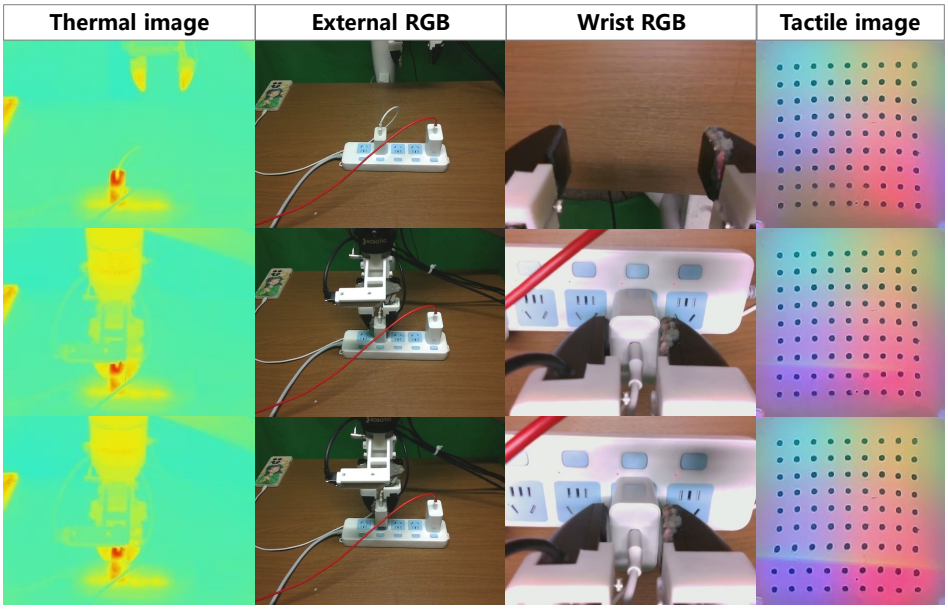

Figure 8: Multimodal observations in the powered-charger disconnection task. Before grasping, the thermal image reveals the heat associated with electrical activity and identifies the powered charger. After grasping, changes in the visuotactile image and tracked-marker positions characterize the physical interaction.

After the target object is grasped, visible deformation appears in the visuotactile image, accompanied by clear displacement of the tracked markers from their no-contact positions. These contact-dependent changes provide the low-level VTLA controller with information about physical contact and grasp stability. The visualizations therefore illustrate the complementary temporal roles of the two physical sensing modalities: thermal sensing resolves the target before interaction, while tactile sensing becomes informative after contact.

D Additional Experimental Analysis

PGA measures the percentage of episodes in which the high-level VLM selects the correct executable prompt, while TSR measures the percentage of real-robot trials that successfully complete the task.

D.1 Contact-dependent marker displacement

Figure 9 compares the tracked tactile-marker displacements before and after contact in the beverage and charger tasks. Before contact, the markers remain close to their reference positions, with only small variations caused by sensor noise and robot motion. Once the gripper establishes physical contact with the Coke can, bread, or charger plug, the marker field exhibits substantially larger and spatially structured displacement. This consistent transition from near-zero displacement to coherent deformation provides a reliable observable cue for determining whether tactile information corresponds to valid physical contact. These qualitative results support the design of CAG, which suppresses tactile features in the no-contact stage and activates them only after the marker displacement exceeds the contact threshold. Together with the quantitative ablation results in the main paper, this analysis indicates that contact-conditioned tactile activation prevents non-contact sensor variation from unnecessarily influencing the action representation.

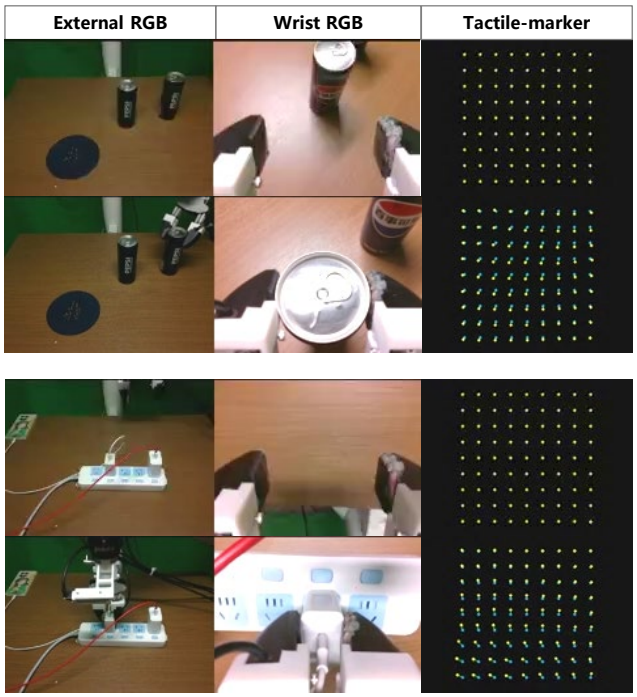

Figure 9: Tactile-marker displacement before and after contact in the beverage, and charger tasks.

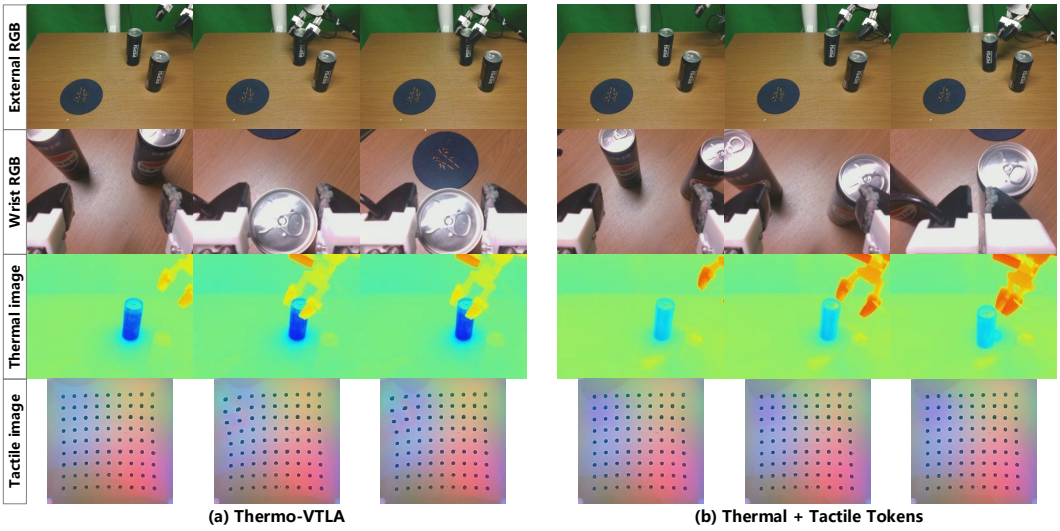

Figure 10: Representative failure of direct thermal-token injection (cold-beverage retrieval).

D.2 Failure of direct thermal-token injection

While the main paper presents a qualitative comparison of direct thermal-token injection in the powered-charger disconnection task, Fig. 10 and Fig. 11 provide the corresponding additional comparisons for the other two tasks: cold-beverage retrieval and room-temperature bread manipulation. In both tasks, although the thermal observations contain clear physical-state differences between the candidates, the direct-injection policy fails to reliably convert this evidence into target-specific robot motion. Importantly, this failure does not indicate that the thermal observation lacks useful physical information. Instead, it suggests that the direct-injection policy does not effectively exploit this information under limited robot demonstrations.

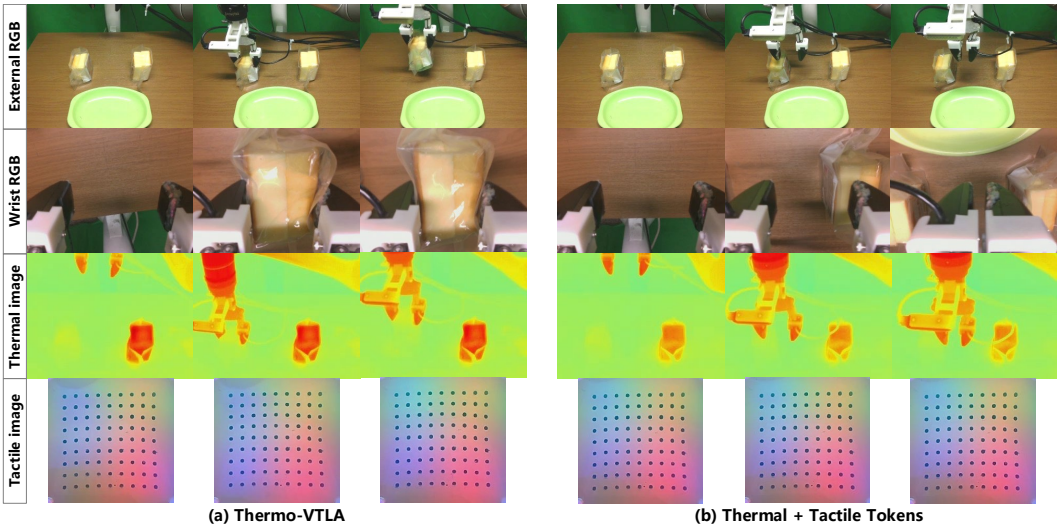

Figure 11: Representative failure of direct thermal-token injection (room-temperature bread manipulation).

A plausible explanation is that the VLA backbone is pretrained primarily on natural RGB–language data and therefore does not possess an explicit thermal–language alignment prior. Direct thermal conditioning consequently requires the policy to simultaneously learn a thermal representation, align it with RGB and language, associate it with the correct spatial target, and generate continuous actions using only sparse action supervision. With a small demonstration dataset, the familiar RGB features can remain dominant, while the out-of-distribution thermal tokens may be underutilized or introduce conflicting features into the action context. When the two candidates have similar RGB appearances, this unresolved target uncertainty can result in an intermediate motion between them.

D.3 Failure without contact-aware tactile integration

Figure 12 further compares stable and unstable grasping cases using visual and tactile observations. From the external and wrist RGB views, the gripper appears to have contacted the object in both cases, making it difficult to determine whether sufficient and stable contact has actually been established. Visual observations alone also provide limited evidence of small relative motion or incipient slip inside the gripper. In contrast, the visuotactile image and marker- displacement field exhibit clearly different contact patterns: a stable grasp produces persistent and structured deformation, whereas an unstable grasp produces weak, asymmetric, or rapidly changing tactile responses.

Without contact-aware tactile integration, the controller cannot reliably distinguish no contact, marginal contact, and stable grasp establishment. It may therefore begin lifting or pulling before the object is securely held. This failure is particularly evident in the charger task, where an insufficient grasp can cause the plug to slip relative to the fingers during the pulling motion, resulting in unsuccessful disconnection. By combining CAG and GF, the full model first determines whether the tactile observation corresponds to valid contact and then adaptively injects the valid tactile representation into the action context. Through the expert demonstrations, the controller learns tactile patterns associated with stable interaction for different objects, allowing it to maintain contact more reliably during lifting, transport, and plug removal. This comparison explains why the benefit of contact-aware tactile integration cannot be fully reproduced by RGB observations alone.

E Discussion, Limitations, and Future Work

Discussion. Thermo-VTLA separates non-contact physical-state grounding from contact-rich visuotactile control. Instead of requiring a single policy to jointly learn thermal–language alignment, tactile-contact modeling, and continuous action generation from limited robot demonstrations, the framework converts thermal evidence into a target-resolved executable prompt before manipulation.

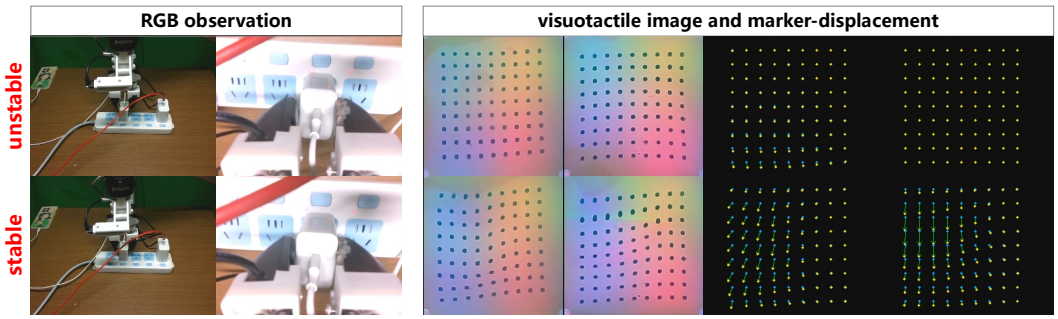

Figure 12: Comparison between unstable and stable grasp establishment. The external and wrist RGB observations appear similar in the two cases, while the visuotactile image and marker-displacement field reveal clear differences in contact deformation and grasp stability. Contact-aware tactile integration enables the controller to distinguish marginal contact from stable object engagement.

The low-level controller can therefore focus on visuomotor execution and contact-dependent tactile feedback. This role-aware decomposition reduces the multimodal learning burden placed on robot demonstrations while preserving the complementary temporal roles of thermal and tactile sensing. It also provides a modular form of task adaptation: new physical-state instructions can be introduced by replacing the candidate prompt set $\mathcal{C}$, without retraining or fine-tuning the high-level reasoner.

An additional benefit is that raw thermal features are not directly provided to the low-level policy. Changes in ambient temperature, object temperature, or thermal-camera appearance between training and deployment therefore do not directly alter the controller input distribution. Once the relevant physical state has been grounded into a discrete spatial prompt, the VTLA controller operates on the same RGB, proprioceptive, and tactile inputs used during training. This design improves modularity and reduces sensitivity to moderate thermal-domain shifts. Nevertheless, the reliability of the overall system still depends on whether the high-level reasoner can correctly interpret the thermal observation.

Limitations of thermal grounding. The current framework queries the thermal-grounded reasoner once at the beginning of each episode. This design assumes that the physical predicate is primarily required for initial target selection and remains sufficiently stable during subsequent execution. Consequently, the current framework is not directly applicable to tasks in which temperature changes dynamically or must continuously influence the robot trajectory, such as heat-aware grasp adjustment, temperature-dependent force control, or manipulation near evolving thermal hazards. Future work could introduce periodic or event-triggered re-grounding, allowing the high-level reasoner to update the executable prompt when a meaningful physical-state change is detected. Another direction is to provide the controller with a compact and calibrated thermal-state estimate, rather than raw thermal tokens, when continuous thermal feedback is necessary.

Limitations of constrained prompt selection. The constrained candidate set improves output consistency and prevents the reasoner from generating commands outside the action capabilities of the trained controller. However, it does not guarantee that the selected prompt is physically correct. Grounding errors may still occur under weak thermal contrast, occlusion, reflective surfaces, sensor noise, or scenes containing many visually similar candidates. Moreover, the current candidate sets are manually specified and contain a small number of spatially resolved commands. Scaling to cluttered environments with more objects, spatial relations, and manipulation skills may require larger or hierarchically structured prompt spaces. Future work could automatically construct $\mathcal{C}$ from detected objects and available robot skills, and represent executable prompts using structured fields such as target, location, and action.

Evaluation scope and future work. The current real-robot evaluation covers three representative tasks involving temperature- and electrical-state grounding, but it is conducted using a single robotic platform and relatively structured laboratory scenes. Future evaluations will include more diverse

ambient conditions, object temperatures, scene layouts, and long-horizon tasks, together with cross-robot and cross-sensor transfer.