MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

TacSushi: Tactile-Grounded World–Action Modeling for Dexterous Food Manipulation

Hu, Haodi; Kogashi, Kaen; Koike-Akino, Toshiaki

TR2026-173 November 13, 2026

Abstract

Dexterous food manipulation requires control under deformation, occlusion, and uncertain contact. We present TacSushi, a Cosmos3-based policy that gives current fingertip touch a feature-wise gated path to the deployed action head while using logged future visual, tactile, progress, and contact outcomes only as training supervision. Failed demonstrations retain consequence targets, but their actions are masked from imitation. Across 600 real-robot rollouts, the full model reaches 68.3% average success on three in-distribution tasks and 37.5% on two held-out ingredients under a per-rollout human–VLM visual-quality rule. Matched ablations support complementary benefits from gated tactile fusion and training-only consequence prediction, while also revealing limited ingredient-level transfer and open questions about which predicted outcome matters most

Conference on Robot Learning (CoRL) Workshop on Dexterous Manipulation: Full-Stack Robotics on Learning and Design 2026

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

TacSushi: Tactile-Grounded World–Action Modeling for Dexterous Food Manipulation

Haodi Hu[1,2†], **Kaen Kogashi**[2,3†], **Toshiaki Koike-Akino**[2]
[†] equal contribution
[1]University of Southern California, USA
[2]Mitsubishi Electric Research Laboratories (MERL), USA
[3]Mitsubishi Electric, Japan

Abstract: Dexterous food manipulation requires control under deformation, occlusion, and uncertain contact. We present TacSushi, a Cosmos3-based policy that gives current fingertip touch a feature-wise gated path to the deployed action head while using logged future visual, tactile, progress, and contact outcomes only as training supervision. Failed demonstrations retain consequence targets, but their actions are masked from imitation. Across 600 real-robot rollouts, the full model reaches 68.3% average success on three in-distribution tasks and 37.5% on two held-out ingredients under a per-rollout human–VLM visual-quality rule. Matched ablations support complementary benefits from gated tactile fusion and training-only consequence prediction, while also revealing limited ingredient-level transfer and open questions about which predicted outcome matters most.

Keywords: dexterous manipulation, tactile sensing, world models

1 Introduction

Fingertip touch supplies local contact evidence when the hand occludes the workspace and external vision cannot resolve force or incipient slip. Food makes this mismatch unusually visible. Sticky rice adheres and deforms, seaweed bends and slips, and toppings can be displaced or damaged by small errors. Ingredient properties also vary between samples [1], while haptics can improve food acquisition [2]. These properties make sushi preparation a compact test bed for contact-sensitive dexterity rather than merely visual task completion.

Prior work has shown that touch supports grasp outcome prediction and regrasp selection [3], shared visual–tactile representations [4], and predictive control [5, 6]. Recent imitation policies use current visual and tactile observations [7] or predict future touch as an auxiliary objective [8]. Yet three questions remain entangled: whether touch should modify the deployed action representation directly, whether future consequences should be used online or only to shape learning, and whether failed interactions can teach without being imitated.

We study these questions through TacSushi, a tactile-grounded world–action policy for a four-finger Shadow Hand. We test two design choices. First, current touch enters the deployed policy through a feature-wise residual gate rather than direct concatenation. Second, recorded future consequences supervise the shared representation during training, but predicted futures are not queried during execution. A matched four-variant comparison holds current tactile access, data, optimization, and the executor fixed while varying these two factors. Appendix A provides implementation and evaluation details.

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.

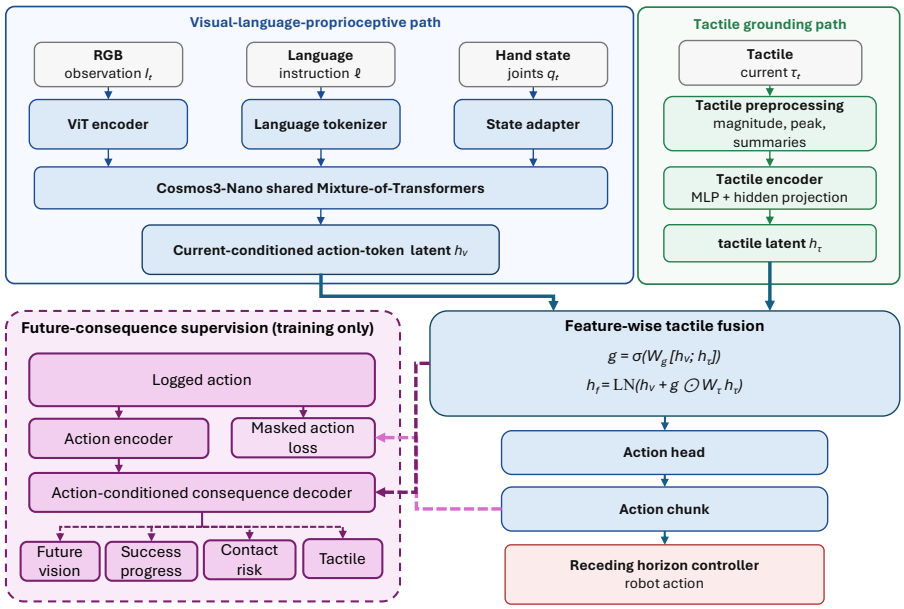


Figure 1: TacSushi gates current touch into the deployed action path (blue). The training-only decoder (purple) predicts demonstrated-action consequences and is removed at deployment.

2 Tactile-Grounded World–Action Modeling

Platform and data. The four-finger Shadow Dexterous Hand Lite has four tactile fingertips (68 three-axis taxels, 204 raw channels). Bare-hand teleoperation retargets MediaPipe landmarks to 13 actuator commands. Two workspace cameras, 16 observed joints, actuator targets, and touch are logged at 30 Hz. Training uses 340 successful trials (114 rice-ball, 113 seaweed, 113 surf-clam) and 50 failures. The policy predicts 16-step chunks of 13-D targets and executes 12 steps before re-observing.

Deployed action path. At time t, a Cosmos3-Nano backbone [9] encodes RGB, language, and the observed hand state into h_v. A compact encoder maps the current tactile measurement to h_τ. Touch enters as a gated residual,

$$g_t = \sigma(W_g[h_v; h_\tau]), \qquad h_f = \mathrm{LN}(h_v + g_t \odot W_\tau h_\tau), \tag{1}$$

and an action head maps h_f to the actuator-target chunk. This direct path is the only mechanism used for deployed control.

Training-only consequence path. During training, a separate decoder receives h_f and an embedding of the demonstrated action chunk and current hand state. It predicts aligned future RGB, task progress, relative contact risk, and a tactile summary. Its loss is

$$\mathcal{L} = m_t \mathcal{L}_{\text{action}} + \lambda_{\text{vis}} \mathcal{L}_{\text{future-vis}} + \mathcal{L}_{\text{aux}}, \tag{2}$$

where $m_t = 1$ for successful demonstrations and 0 for failures. Thus a failed window conditions the decoder on the command that produced its logged adverse outcome, but that command is not an imitation target. The auxiliary loss combines progress, risk, tactile-summary, and consistency terms (Appendix A.1). Its gradients shape the action representation; the action head never consumes predicted consequences. The action encoder and decoder are removed at deployment.

Controlled variants. M3/M4 concatenate touch without/with consequence supervision; M5/M6 use the gate without/with it. M6 is full TacSushi. These four share current tactile access, data order, training steps, optimizer, action loss, and executor. M1 is GR00T [10] without touch; M2 adds gated touch. The matched M3–M6 pairs test the two design choices; M1–M2 provide system comparisons.



Figure 2: Representative successful executions. The top three rows are in-distribution tasks; black rice and salmon are held-out ingredient variants. The montage shows task evolution that a terminal score omits.

3 Real-Robot Evidence

Protocol. We evaluate rice-ball shaping, seaweed wrapping, and surf-clam placement, plus held-out black-rice and salmon variants. No held-out ingredient is used for task-specific adaptation. Each method–task cell contains 20 trials, yielding 600 evaluation rollouts separate from training. Acceptable outcomes differ by task, so a fixed geometric threshold would conflate natural food variation with failure. Each terminal image instead receives five human ratings and one rating from each of three vision–language models on an anchored five-point visual-quality scale. The eight ratings are averaged with equal weight; a trial succeeds when its composite score is at least 3 (acceptable). This measures visible completion and damage. Appendix A.3 reports the survey protocol and separate human and AI scores.

Main result. Full TacSushi reaches 75%, 60%, and 70% success on the three in-distribution tasks (68.3% average), and 35% and 40% on black rice and salmon (37.5% average). Its descriptive five-task pooled rate is 56%, versus 31% for the strongest system baseline, GR00T+tactile, a 25-point gap. Figure 3 reports all conditions with Wilson intervals.

What the controlled contrasts show. At fixed direct-concatenation fusion, consequence supervision raises average in-distribution/out-of-distribution (ID/OOD) success from 15.0/5.0% (M3) to 25.0/17.5% (M4). At fixed gated fusion, it raises success from 36.7/10.0% (M5) to 68.3/37.5% (M6). Conversely, replacing concatenation by the gate raises ID/OOD success from 15.0/5.0% to 36.7/10.0% without consequence supervision, and from 25.0/17.5% to 68.3/37.5% with it. Because M5 and M6 have identical deployed inputs and architecture, their 31.7-point ID and 27.5-point OOD gaps support a benefit from training-only predictive supervision without extra deployment-time sensing or decoder computation.

An offline diagnostic supplies complementary evidence. On 176 trajectory-disjoint clips (5,808 frames per method), future-vision errors for the gated models fall from 5.6% MAE, 22.9% $1-\text{SSIM}$, and 0.17 LPIPS (M5) to 3.8%, 17.7%, and 0.13 (M6). The decoder and shared representation are both updated, so this test does not isolate latent quality.

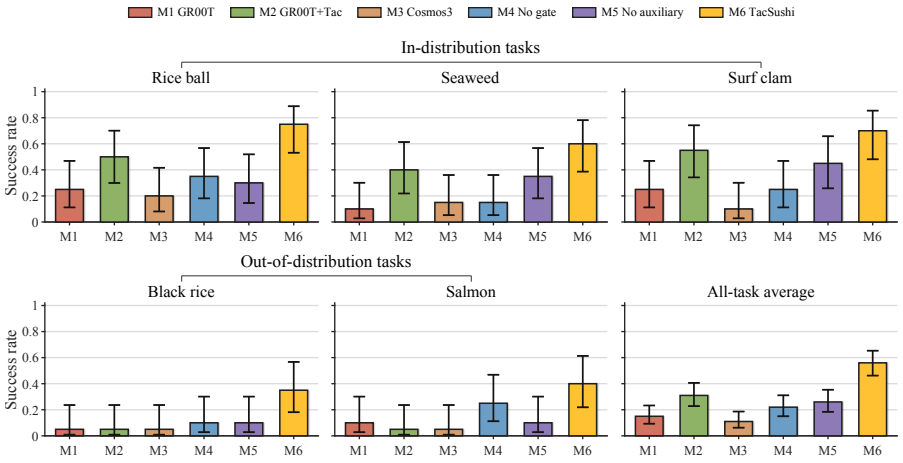


Figure 3: Success under the per-rollout human–VLM outcome rule. Error bars are 95% Wilson intervals with $n = 20$ per task and $n = 100$ for the descriptive pooled panel. M3–M6 are the matched variants.

4 Limitations and Future Directions

Our controlled comparisons demonstrate the complementary benefits of tactile-gated fusion and training-time consequence supervision, but do not isolate the contributions of individual predictive objectives. In particular, removing all consequence losses simultaneously cannot distinguish the effects of future-vision prediction, tactile summaries, progress and risk supervision, or learning from failed trajectories. Future work should isolate these components through targeted ablations, particularly by excluding failed-window supervision while retaining the remaining objectives. Moreover, the consequence decoder serves as a training-time auxiliary objective rather than a validated causal dynamics model.

Our contact-risk and progress targets also have limitations. Normalization by each episode's contact quantile captures relative contact intensity but discards absolute differences across episodes. Consequently, even relatively gentle interactions may receive high risk labels. Clipping the future-window peak risk at one can further suppress the progress target to zero during otherwise successful manipulation. These heuristic targets do not necessarily distinguish harmful contact from the pressure required for food shaping. Analyzing target distributions and saturation frequencies across successful and failed trajectories would help assess their informativeness. Future extensions could incorporate cross-episode contact calibration and contact-aware measures of slip or deformation.

Evaluation introduces additional uncertainty. Our success metric combines five human and three VLM ratings of terminal visual quality. Although VLM ratings provide an additional rubric-based assessment, they are not independent ground truth, and their inclusion may influence decisions near the success threshold. Human-only scoring and threshold-sensitivity analyses would strengthen the robustness of the evaluation. Furthermore, Wilson intervals capture rollout variability for fixed checkpoints and evaluation labels, but not training-seed or evaluator uncertainty. Repeated training runs and explicit modeling of rater disagreement would support stronger statistical conclusions.

Finally, ingredient generalization remains limited. Success decreases from 68.3% on in-distribution tasks to 37.5% on two held-out ingredients. These evaluations preserve the learned manipulation objectives and workspace, and therefore do not establish generalization to new tasks, embodiments, or unstructured environments. Potential failure mechanisms include slipping, adhesion during release, incomplete wrapping, and topping displacement, which terminal visual ratings alone cannot reliably characterize. A promising direction is to adapt the policy using logged successes and failures from unfamiliar ingredients, followed by fresh evaluations of adaptation performance and retention on previously learned tasks.

References

[1] Z. Wang, S. Hirai, and S. Kawamura. Challenges and opportunities in robotic food handling: A review. *Frontiers in Robotics and AI*, 8:789107, 2021. doi:10.3389/frobt.2021.789107.

[2] T. Bhattacharjee, G. Lee, H. Song, and S. S. Srinivasa. Towards robotic feeding: Role of haptics in fork-based food manipulation. *IEEE Robotics and Automation Letters*, 4(2):1485–1492, 2019. doi:10.1109/LRA.2019.2894592.

[3] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018. doi:10.1109/LRA.2018.2852779.

[4] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *IEEE International Conference on Robotics and Automation*, 2019.

[5] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine. Manipulation by feel: Touch-based control with deep predictive models. In *IEEE International Conference on Robotics and Automation*, pages 818–824, 2019. doi:10.1109/ICRA.2019.8794219.

[6] W. Mandil, K. Nazari, and A. Ghalamzan. Action conditioned tactile prediction: Case study on slip prediction. In *Proceedings of Robotics: Science and Systems*, 2022. doi:10.15607/RSS.2022.XVIII.070.

[7] M. Murooka, T. Hoshi, K. Fukumitsu, S. Masuda, M. Hamze, T. Sasaki, M. Morisawa, and E. Yoshida. TACT: Humanoid whole-body contact manipulation through deep imitation learning with tactile modality. *IEEE Robotics and Automation Letters*, 10(8):7819–7826, 2025. doi:10.1109/LRA.2025.3580329.

[8] L. Heng, H. Geng, K. Zhang, P. Abbeel, and J. Malik. ViTacFormer: Learning cross-modal representation for visuo-tactile dexterous manipulation. *arXiv preprint arXiv:2506.15953*, 2025.

[9] NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL `https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf`.

[10] NVIDIA. Isaac GR00T N1.7: A foundation model for generalist robots. `https://github.com/NVIDIA/Isaac-GR00T`, 2026. Accessed: 2026-06-17.

[11] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann. MediaPipe Hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi:10.1109/TIP.2003.819861.

[13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 2018.

[14] X. Lin, Y. Wang, J. Olkin, and D. Held. SoftGym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 432–448. PMLR, 2021.

[15] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu. RoboCraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks. *The International Journal of Robotics Research*, 43(4):533–549, 2024. doi:10.1177/02783649231219020.

[16] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[17] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.

[18] C. Finn and S. Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation*, pages 2786–2793, 2017. doi:10.1109/ICRA.2017.7989324.

A Implementation and Additional Evidence

This appendix provides implementation, target construction, evaluation protocol, detailed success rates, human–AI comparison, diagnostic, and limitations that support the four-page workshop narrative.

A.1 System, Data, and Training Targets

Platform and collection. The four-finger Shadow Dexterous Hand Lite has 13 actuators and 16 observed joints, including three passive joints that are observed but not commanded. Four Shadow Tactile Fingertips provide 17 three-axis taxels each (68 taxels; 204 raw scalar channels). Per-taxel Euclidean magnitudes give 68 tactile features; current-input summaries use samples available at or before t. A RealSense/MediaPipe interface [11] retargets bare-hand landmarks using fixed gains and offsets to actuator commands at up to 10 Hz; targets are clipped, low-pass filtered ($\alpha = 0.30$), limited to 0.10 rad per step, interpolated, and checked for self-collision. Two workspace cameras, robot state/targets, and touch are logged at 30 Hz; the operator-view stream is not a policy observation. Figure 4 shows the workspace and collection interface.

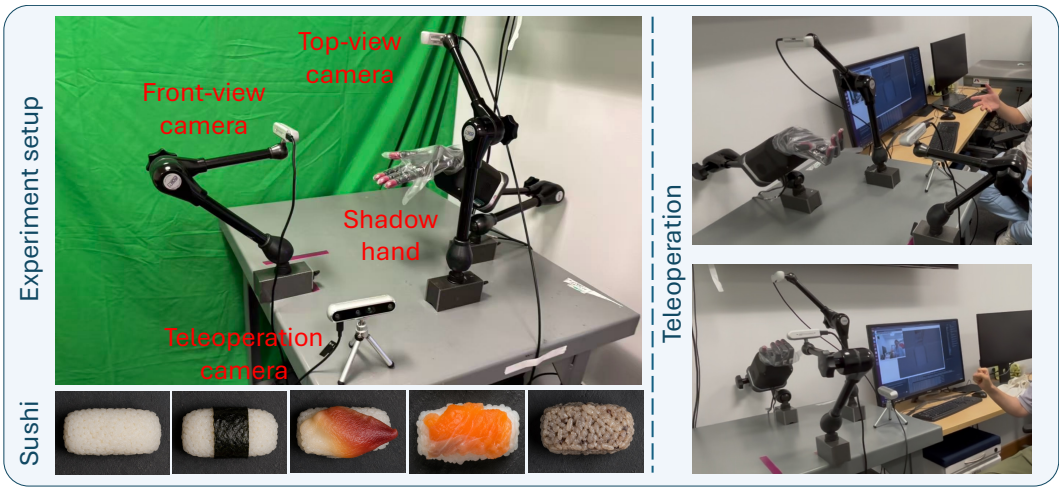


Figure 4: Physical platform and bare-hand teleoperation. Front and top cameras record the workspace; a separate operator-view camera supports MediaPipe tracking. Workspace RGB, robot state/targets, and fingertip touch are logged synchronously.

Table 1: Training corpus. Durations and frames per workspace-camera stream are averaged over successful trials.

Statistic	Rice ball	Seaweed	Surf clam	Overall
Successful trials	114	113	113	340
Failed trials	16	17	17	50
Avg. duration (s)	74	117	71	87.3
Avg. RGB frames	2,220	3,510	2,130	2,619

Aligned consequence supervision. Each window contains current RGB, language, 16-joint state, and touch; a demonstrated 16×13 action chunk; 33 aligned future RGB frames; and progress, contact-risk, and tactile-summary targets. Actions are resampled to 16 steps and RGB to 33 frames over the same physical interval with aligned endpoints. The decoder receives h_f and $E_a(\mathbf{a}_t^\star, q_t)$, an embedding of the demonstrated chunk and current joint state; its losses back-propagate into the shared representation. The 33-sample tactile horizon spans 1.10 s at 30 Hz. Let $\boldsymbol{\tau}_s^{\mathrm{mag}} \in \mathbb{R}^{68}$ be tactile magnitudes, $u_s = \max_k \tau_{s,k}^{\mathrm{mag}}$, $\mathcal{S}_t = \{t+1, \ldots, t+33\}$, $\rho_t = \min\{(t+33)/T, 1\}$, and m_t

indicate episode success. With the episode's 0.90 contact quantile $Q_{.90}$,

$$y_t^{\text{risk}} = \text{clip}\left(\frac{\max_{s\in\mathcal{S}_t} u_s}{Q_{.90}(u_{1:T}) + 10^{-6}}, 0, 1\right), \qquad y_t^{\text{value}} = m_t\rho_t(1 - y_t^{\text{risk}}),$$
$$b_t^{\text{tac}} = [\boldsymbol{\tau}_{t+33}^{\text{mag}}; \tfrac{1}{33}\sum_{s\in\mathcal{S}_t} \boldsymbol{\tau}_s^{\text{mag}}; y_t^{\text{risk}}; \rho_t] \in \mathbb{R}^{138}. \tag{3}$$

The risk target is an episode-relative peak-contact proxy, rather than a calibrated probability of slip or damage. Progress is successful episode phase discounted by this proxy; the tactile summary combines endpoint and mean contact with peak risk and phase. Future samples, the full-episode quantile, and episode outcome are privileged training targets and never enter the deployed action head. The training loss is

$$\begin{aligned}\mathcal{L} = m_t\mathcal{L}_{\text{action}} + \lambda_{\text{vis}}\mathcal{L}_{\text{future-vis}} + \lambda_v(\hat{v} - y^{\text{value}})^2 + \lambda_r(\hat{r} - y^{\text{risk}})^2 \\ + \frac{\lambda_\tau}{138}\|\hat{b} - b^{\text{tac}}\|_2^2 + \lambda_c(\hat{r} - [\hat{b}]_{\text{risk}})^2.\end{aligned} \tag{4}$$

Only imitation is masked on failed trials; their logged adverse consequences remain supervised. Selected backbone adapters and task-specific action, tactile, fusion, and consequence modules are optimized jointly while most pretrained parameters remain frozen. The loss weights are fixed and shared across the controlled variants; the final term aligns the two risk readouts. At deployment, the decoder and demonstrated-action encoder are removed; the controller clamps targets, checks self-collision, executes 12 of 16 predicted steps, and re-observes.

A.2 Evaluation Protocol and Controlled Results

Rice-ball shaping compacts rice into a self-supporting oval; seaweed wrapping closes a sheet around the rice; surf-clam placement centers and retains the topping without visible crushing. Black rice preserves shaping with a new ingredient, and salmon preserves placement with a new topping. No held-out ingredient is used for task-specific adaptation. M1 is GR00T without touch; M2 adds gated current touch. M3–M6 share data, order, steps, optimizer, action loss, and executor: M3 is concatenate/no auxiliary loss, M4 concatenate/auxiliary, M5 gate/no auxiliary, and M6 gate/auxiliary. Disabling the auxiliary objectives sets all consequence-loss weights to zero while preserving current touch. Each method–task cell has 20 fresh rollouts. Table 2 gives the complete 600-rollout result; none of these rollouts is used for updates.

Table 2: Success rates (%). ID Avg. covers rice ball (RB), seaweed (SW), and surf clam (SC); OOD Avg. covers black rice (BR) and salmon (SA).

Method	RB	SW	SC	ID Avg.	BR	SA	OOD Avg.
M1 GR00T	25	10	25	20.0	5	10	7.5
M2 GR00T+tactile	50	40	55	48.3	5	5	5.0
M3 concatenate, no aux.	20	15	10	15.0	5	5	5.0
M4 concatenate + aux.	35	15	25	25.0	10	25	17.5
M5 gate, no aux.	30	35	45	36.7	10	10	10.0
M6 TacSushi	**75**	**60**	**70**	**68.3**	**35**	**40**	**37.5**

Future-vision diagnostic. On 176 trajectory-disjoint clips (33 targets each; 5,808 frames per method), consequence-supervised models improve every reconstruction metric (Table 3). M3 and M5 retain the pretrained visual readout without future-vision adaptation; M4 and M6 adapt it during training. RGB is normalized to $[0, 1]$. Metrics are $100\times\text{MAE}$, frame-averaged $100\times(1-\text{SSIM})$ with an 11×11 Gaussian window [12], and LPIPS [13]. Figure 5 shows qualitative predictions across all five categories. Coarse scene evolution is preserved, with visible errors around rice and seaweed. Because training updates both the shared representation and decoder, this diagnostic does not isolate latent quality.

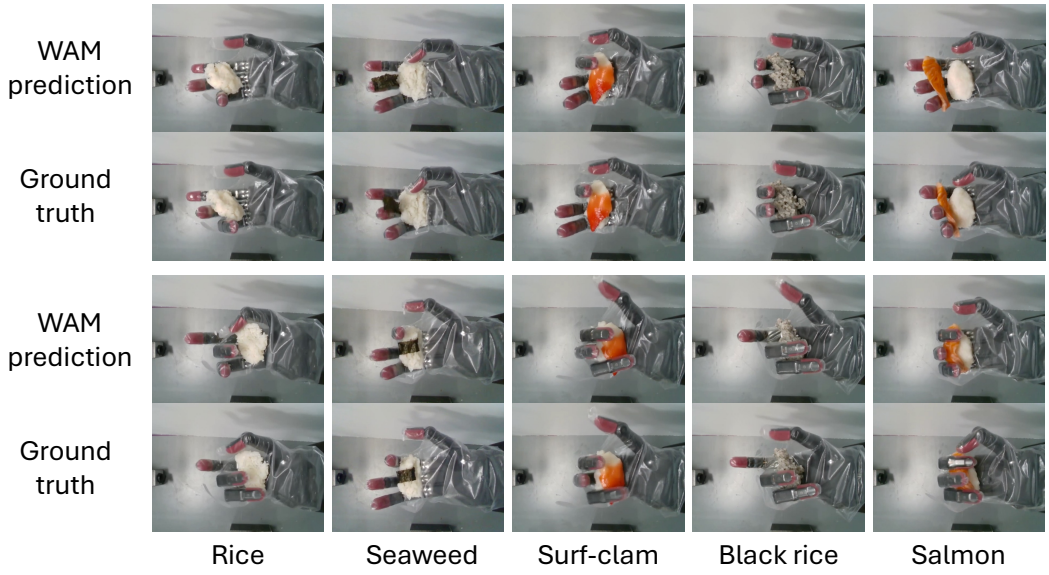


Figure 5: Full TacSushi future-vision predictions paired with synchronized ground truth for all five food categories. Each pair places prediction above ground truth. The visual readout is used only for training and offline diagnostics.

Table 3: Offline future-vision reconstruction; lower is better.

Model	MAE (%)	$1 - \text{SSIM}$ (%)	LPIPS
M3	11.4	35.1	0.29
M4	7.5	28.9	0.21
M5	5.6	22.9	0.17
M6	**3.8**	**17.7**	**0.13**

A.3 Human–VLM Food-Outcome Evaluation

Fifty voluntary, uncompensated participants provided written consent. Each participant rated 60 randomly assigned terminal images (12 per category), and each of the 600 images received five human ratings, giving 3,000 assignments. The anchored scale was 5 excellent, 4 good, 3 acceptable, 2 poor, and 1 very poor. Criteria emphasized an oval, self-supporting rice portion, closed wrap, balanced topping placement, gentle pressing, and avoidance of crushing. The study underwent institutional operations review (institution withheld for double-blind review), took approximately 15 minutes per participant, and anonymized responses. Participants were told that images could depict chef, conveyor-belt, or robotic-hand preparation; individual images did not display source labels.

GPT-5.6, Claude Opus 4.8, and Gemini 3.5 independently rated the same images under a common prompt with the human survey's instructions and rubric. For rollout j, $\mathcal{H}_j$ is the set of five human raters and the three AI ratings are indexed by k:

$$s_j = \frac{1}{8}\left(\sum_{h\in\mathcal{H}_j} r_{j,h}^{\text{human}} + \sum_{k=1}^{3} r_{j,k}^{\text{AI}}\right), \qquad \text{success}_j = \mathbf{1}[s_j \geq 3]. \tag{5}$$

Thus every rating has equal weight, while the human group contributes $5/8$. The composite is computed separately for each rollout. Ratings assess visible completion and damage, not taste, texture, or food safety. Because they define success, they are not an independent validation set. Wilson intervals use $n = 20$ per task and $n = 100$ only for the descriptive five-task pool.

Table 4: Human and AI-ensemble visual-quality scores (mean ± SEM; 20 rollouts per method–category cell). Each mean and SEM summarizes the corresponding rollout-level evaluator means. Category MAE is recomputed from the six displayed human–AI mean pairs, in rating points.

Category	Method	Human	AI ensemble	MAE (↓)
Rice ball	M1	1.96 ± 0.29	2.02 ± 0.26	0.07
	M2	2.66 ± 0.32	2.65 ± 0.35	
	M3	1.80 ± 0.26	1.91 ± 0.28	
	M4	2.22 ± 0.31	2.28 ± 0.27	
	M5	2.02 ± 0.29	2.18 ± 0.31	
	M6	$\mathbf{3.34 \pm 0.28}$	$\mathbf{3.31 \pm 0.26}$	
Seaweed	M1	1.42 ± 0.19	1.62 ± 0.18	0.10
	M2	2.26 ± 0.31	2.42 ± 0.29	
	M3	1.68 ± 0.23	1.77 ± 0.22	
	M4	1.68 ± 0.23	1.76 ± 0.22	
	M5	2.22 ± 0.31	2.28 ± 0.29	
	M6	$\mathbf{2.92 \pm 0.31}$	$\mathbf{2.93 \pm 0.34}$	
Surf clam	M1	1.92 ± 0.28	2.04 ± 0.24	0.10
	M2	2.78 ± 0.32	2.76 ± 0.30	
	M3	1.46 ± 0.19	1.62 ± 0.18	
	M4	1.88 ± 0.28	2.06 ± 0.26	
	M5	2.50 ± 0.28	2.58 ± 0.30	
	M6	$\mathbf{3.20 \pm 0.29}$	$\mathbf{3.24 \pm 0.28}$	
Salmon (OOD)	M1	1.44 ± 0.17	1.62 ± 0.18	0.13
	M2	1.58 ± 0.23	1.77 ± 0.22	
	M3	1.38 ± 0.14	1.47 ± 0.13	
	M4	1.98 ± 0.24	2.10 ± 0.26	
	M5	1.52 ± 0.19	1.58 ± 0.18	
	M6	$\mathbf{2.26 \pm 0.34}$	$\mathbf{2.42 \pm 0.30}$	
Black rice (OOD)	M1	1.30 ± 0.16	1.52 ± 0.13	0.11
	M2	1.58 ± 0.23	1.78 ± 0.22	
	M3	1.36 ± 0.14	1.47 ± 0.13	
	M4	1.52 ± 0.22	1.50 ± 0.18	
	M5	1.54 ± 0.19	1.56 ± 0.18	
	M6	$\mathbf{2.27 \pm 0.29}$	$\mathbf{2.18 \pm 0.29}$	

M6 has the highest mean from both evaluator groups in every category. The lower held-out scores nevertheless show limited ingredient-level transfer. The category MAEs describe aggregate score agreement, not image-level rater agreement or statistical equivalence. There are no chef-made or conveyor-belt reference samples in this comparison.

A.4 Scope, Limitations, and Adaptation

The study uses one training run and checkpoint per method. Wilson intervals condition on fixed models and labels; score inference should account for participants rating multiple images and ratings sharing an image. Ingredient shifts preserve the task and embodiment. Individual consequence targets and failed-window supervision remain unablated, and terminal images do not measure failure-mechanism frequencies. A chronological adaptation study should update from a new ingredient's logged outcomes, mask failed commands, and evaluate fresh rollouts and retention on the original foods. Common frozen probes and target-specific ablations would test which outcomes shape the action features.

Relation to the broader literature. Deformable-material learning [14, 15], generalist policies [16], action chunking [17], and online visual foresight [18] provide complementary precedents. TacSushi contributes the coupling of a deployed tactile gate with training-only action-consequence targets and its controlled evaluation.