Supplementary Figures and Tables: When Replanning Becomes the Bottleneck: Budgeted Replanning for Embodied Agents

Anonymous Project Page: https://anonymous-2026.github.io/BRACE-ICML

Table 1. RoboFactory budget-matched selector-side and recent-method comparisons on the Pass-Shoe anchor ($B = 128$). All methods in this table solve the task successfully, so the Success column is omitted and we report successful-task cost metrics only.

Method	Lat P95	SLO viol.	Tok after	Wait	Type
E-RECAP	207.09	0.49%	**125.66**	**5776.50**	learned pruning
kvtc (Staniszewski and Łańcucki, ICLR'2026)	210.33	1.01%	125.86	5979.35	quantization
Robust Compression Boundary (Yu et al., IJCAI'2025)	210.33	4.0%	125.86	5979.35	compression
TurboQuant (Zandieh et al., ICLR'2026)	221.57	1.73%	125.90	6021.51	quantization
Cross Distillation Compression (Wang et al., KDD'2025)	221.57	7.0%	125.90	6021.51	distillation
ReST-KV (An et al., ICLR'2026)	231.95	2.27%	125.81	6807.52	selector
Sub-LIME Rank-Correlation Selection (Saranathan et al., ACL'2025)	231.95	9.0%	125.81	6807.52	selector
DefensiveKV (Feng et al., ICLR'2026)	239.09	3.49%	125.83	6862.33	selector
Gated Attention (Qiu et al., NeurIPS'2025)	239.10	14.0%	125.83	6862.33	selector
Grad-Hidden Saliency	239.67	2.43%	125.98	7079.05	derived selector
FreeKV (Liu et al., ICLR'2026)	239.99	3.47%	126.64	7269.27	retrieval
Token Recycling (Luo et al., ACL'2025)	239.99	14.0%	126.64	7269.27	recycling

Table 2. RoboFactory budget-matched heuristic baselines on the Pass-Shoe anchor ($B = 128$). This table reports the heuristic-only comparison under the same accounting contract.

Method	Tokens	Lat P95	Lat P99	Bind rate (%)
Random truncation	128	**200.8**	239.6	94.9
Recency truncation	128	204.9	**225.6**	**95.0**
Structured summary	128	314.4	347.3	94.2

Table 3. Focused single-arm real-robot results on PICKFRUIT and PUSHT. This compact view shows that the same budgeting and replanning interface remains beneficial in physical execution.

Task	Method	Succ. Rate	Replans/Ep	P95 Replan (s)	SLO Viol.
PICKFRUIT	One-Shot	8.0%	0.0	N/A	N/A
PICKFRUIT	No BRACE	24.0%	3.4	29.4	42.7%
PICKFRUIT	**BRACE + E-RECAP**	**40.0%**	**1.9**	**22.8**	**18.6%**
PUSHT	One-Shot	0.0%	0.0	N/A	N/A
PUSHT	No BRACE	12.0%	3.8	34.7	61.3%
PUSHT	**BRACE + E-RECAP**	**32.0%**	**2.2**	**26.1**	**27.5%**

Table 4. Habitat phase-overhead summary on the PointNav slice. Pruning adds only tens of milliseconds relative to multi-second planner latency while restoring schedulability.

Variant	Prune mean	Planner mean	End-to-end mean	End-to-end P95	Tok. after
No BRACE	0.76	3804.74	3808.17	5492.08	256.85
Pruning only (no BRACE gate)	37.21	2414.07	2451.51	2488.24	20.21
BRACE + E-RECAP	35.63	2414.05	2449.92	2486.31	20.21
Recency truncation B20	0.69	2287.56	2288.44	2289.53	20.00

Table 5. Real-robot setup, including the robot, camera stack, model inputs, planner/executor choices, and evaluation budget.

Component	Choice
Robot	Songling PiPER, single-arm
Robot control stack	`piper_sdk` over CAN
Cameras	2 × Orbbec RGB-D + 1 record-only overview camera
Model input modality	RGB only
Camera SDK	OrbbecSDK v1
LLM planner	Qwen2.5-14B-Instruct
VLA executor	OpenVLA
Real-robot tasks	PickFruit, PushT
Evaluation budget	25 episodes / method / task / condition
Video categories	Motivation, Replanning in Action, Method Comparison

Table 6. Per-domain setup parameters used throughout the evaluation. The domain tables report budget settings, while the controller-side stability knobs (δ, ω, w) are audited separately in the proxy sweep.

Platform	Ep/var	SLO (ms)	Agents	Budget(s)
Meta Habitat	30	2,500	1	B=20 (budget-match)
RoboFactory	10	250	multi-agent	B=128 (RAG×Prune)
Microsoft AirSim	10	2,500	K=8	N/A

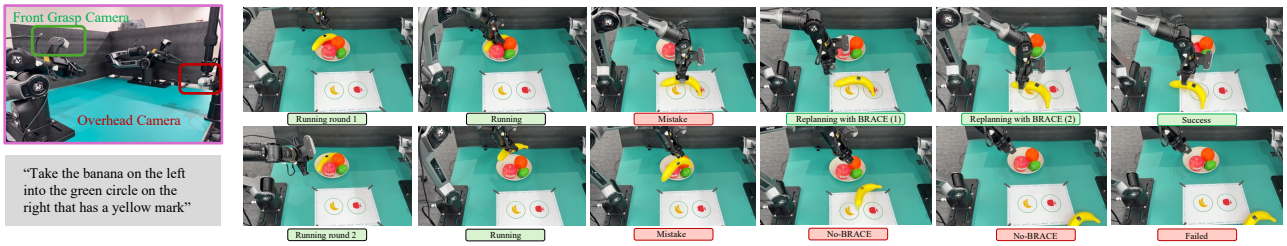

Figure 1. Focused real-robot PICKFRUIT example on a banana instance. The upper rollout uses BRACE and triggers replanning twice after an intermediate mistake, eventually recovering and completing the task; the lower rollout relies only on the underlying LLM+VLA stack and fails after a comparable mistake.

Table 7. Habitat pairwise summary for the no-replanning boundary and budgeted repeated-replanning comparison. '–' denotes metrics not included for a given pair. On this PointNav slice, the principal difference is schedulability rather than task-quality collapse.

Comparison	Task quality	Replans/ep	Wall ms	SLO %	Lat P95 (ms)
No-initial-plan → No BRACE	53.3 / 0.519 → 53.3 / 0.515	0.000 → 4.533	94 → 17451	–	–
No BRACE → BRACE + E-RECAP	53.3 / 0.515 → 53.3 / 0.519	–	–	93.4 → 0.0	5492 → 2486
No BRACE → Recency B20	53.3 / 0.515 → 53.3 / 0.519	–	–	93.4 → 0.0	–

Table 8. Detailed real-robot results for PICKFRUIT and PUSHT, including episode counts, duration, intervention statistics, and task-specific failure categories.

Task	Method	Ep	Succ.	Rate	Dur. (s)	Replans/Ep	P95 Replan (s)	SLO Viol.	Failure A	Failure B
PICKFRUIT	One-Shot	25	2/25	8.0%	402	0.0	N/A	N/A	grasp failure=11	drop/slip=6
PICKFRUIT	No BRACE	25	6/25	24.0%	458	3.4	29.4	42.7%	grasp failure=9	drop/slip=5
PICKFRUIT	**BRACE + E-RECAP**	25	10/25	**40.0%**	**369**	**1.9**	**22.8**	**18.6%**	grasp failure=7	drop/slip=4
PUSHT	One-Shot	25	0/25	0.0%	579	0.0	N/A	N/A	goal error>thr=14	contact loss=12
PUSHT	No BRACE	25	3/25	12.0%	632	3.8	34.7	61.3%	goal error>thr=12	contact loss=15
PUSHT	**BRACE + E-RECAP**	25	8/25	**32.0%**	**501**	**2.2**	**26.1**	**27.5%**	goal error>thr=8	contact loss=7

Table 9. Controller stability sweep as a numeric audit behind the anti-churn discussion. Without failure-aware overrides, larger cooldown and commit windows reduce plan churn but can also hurt success when they become too restrictive.

Variant	δ	ω	w	Success (%)	Calls/ep	Changes/ep	χ	Deadlocks/ep	Stall/ep
No BRACE + E-RECAP	N/A	N/A	N/A	5.0	167.18	166.08	0.993	157.53	225.10
BRACE + E-RECAP	0	0	1	98.3	14.37	13.10	0.912	0.00	91.20
BRACE + E-RECAP	0	0	3	25.0	95.72	94.45	0.987	88.47	189.58
BRACE + E-RECAP	0	0	5	6.7	159.68	158.50	0.993	154.63	222.97
BRACE + E-RECAP	0	2	1	75.0	24.40	23.33	0.956	14.78	131.93
BRACE + E-RECAP	0	2	3	13.3	125.98	124.85	0.991	121.15	211.17
BRACE + E-RECAP	0	2	5	6.7	165.50	164.40	0.993	162.13	219.53
BRACE + E-RECAP	3	0	1	81.7	22.90	21.73	0.949	11.12	129.42
BRACE + E-RECAP	3	0	3	18.3	110.73	109.60	0.990	104.43	197.65
BRACE + E-RECAP	3	0	5	5.0	175.25	174.15	0.994	171.02	227.07
BRACE + E-RECAP	3	2	1	78.3	21.78	20.72	0.951	12.12	125.32
BRACE + E-RECAP	3	2	3	15.0	141.27	140.20	0.992	137.28	207.17
BRACE + E-RECAP	3	2	5	3.3	161.98	160.90	0.993	158.37	225.80

Table 10. Open-loop and harder-setting evidence snapshot. The Habitat rows compare no-replanning and budgeted repeated replanning; the RoboFactory rows compare open-loop, frozen-plan, and BRACE-based recovery under a harder Pass-Shoe setting.

Domain	Method	Task metric	Cost metric	Lat P95	SLO viol.
Habitat	No-initial-plan	53.3% / 0.519 SPL	0 replans	N/A	N/A
Habitat	No BRACE	53.3% / 0.515 SPL	4.533 replans/ep	5492	93.4%
Habitat	BRACE + E-RECAP	53.3% / 0.519 SPL	4.533 replans/ep	2486	0.0%
RoboFactory	Open-loop	0.0% success	0 wait	N/A	N/A
RoboFactory	Frozen plan	0.0% success	55.6 ms wait	72.8	0.0%
RoboFactory	No BRACE	0.0% success	6607.3 ms wait	312.7	27.6%
RoboFactory	BRACE + E-RECAP	80.0% success	6241.7 ms wait	247.2	4.6%

Figure 2. AirSimNH qualitative comparison. The baseline accumulates delayed replanning decisions in the same trigger window, whereas BRACE shortens the effective replanning path and completes the interaction with fewer deadline misses. The corresponding trigger-audit and safety-proxy summaries are reported separately.

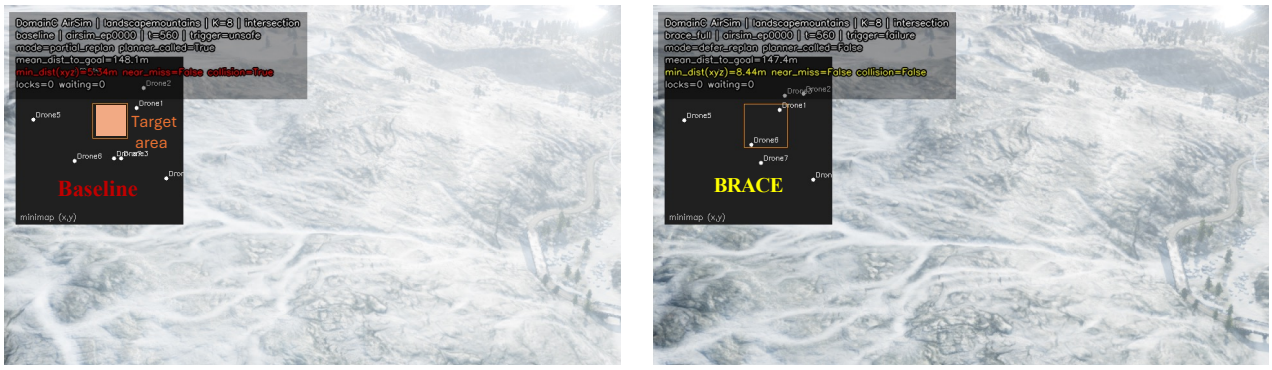

Figure 3. Microsoft AirSim qualitative example in the LandscapeMountains scene (K=8 agents). Left: a baseline that still triggers replanning after collision- or disturbance-induced drift but does not apply BRACE's controller-level budgeting and stabilization. Right: BRACE, which decides whether to honor each trigger, allocates per-call B_t and SLO_t when replanning is admitted, and suppresses replanning churn with cooldown, commit, and failure-aware override. Frames are taken at the same time step and show one drone's first-person view together with the task-region third-person view; the orange overlay marks the goal region. BRACE has already brought agents to completion, while the baseline is still behind.

Figure 4. RoboSuite TWO-ARM PEG-IN-HOLE full-rollout comparison. The upper storyboard shows a BRACE rollout that stabilizes replanning and completes the task; the lower storyboard shows the corresponding no-BRACE baseline, which continues to trigger replanning without controller-level budgeting or stabilization and ultimately fails. Presenting the full contact sheets makes the success-versus-failure contrast directly visible.

Table 11. Simulation platform, scene, and task coverage across the broader evaluation footprint.

Category	Platform	Scene family	Reported task / setting	Role	Status	Where used
Navigation	Meta Habitat	MP3D	PointNav	core navigation benchmark	reported	main text + appendix
Navigation	Meta Habitat	MP3D + shortest-path noise	PointNav with shortest-path noise	stress-test slice	reported	main text + appendix
Manipulation	RoboFactory	indoor handoff	Pass-Shoe	core manipulation benchmark	reported	main text + appendix
Manipulation	RoboFactory	harder handoff	Pass-Shoe harder-setting	harder-setting manipulation slice	reported	main text + appendix
Manipulation	RoboFactory	multi-agent photo task	TakePhoto	qualitative manipulation example	reported	appendix figure
Manipulation	RoboFactory	camera alignment	CameraAlignment	additional manipulation coverage	reported	appendix coverage
Manipulation	RoboSuite	dual-arm assembly	Two-Arm Peg-in-Hole	failure-case extension	reported	main text figure
Manipulation	RoboSuite	dual-arm transfer	Two-Arm Handover	success-case extension	reported	appendix figure
Manipulation	LIBERO	kitchen scene 3	Moka Pot task	cross-benchmark qualitative extension	reported	appendix figure
Manipulation	RoboCasa	kitchen / fruit-style tasks	PickPlaceCounterToCabinet, fruit-style tasks	additional task-family coverage	partial	coverage only
Traffic/UAV	Microsoft AirSim	AirSimNH	intersection / crossing	core traffic/UAV benchmark	reported	main text + appendix
Traffic/UAV	Microsoft AirSim	AbandonedPark	delayed replanning realism bundle	additional AirSim coverage	reported	appendix coverage
Traffic/UAV	Microsoft AirSim	LandscapeMountains	baseline vs BRACE side-by-side ($K=8$)	terrain-shift qualitative example	reported	appendix figure
Traffic/UAV	Isaac Sim	disturbance replay / parking	delayed replanning numeric evidence	additional simulator coverage	partial	appendix coverage

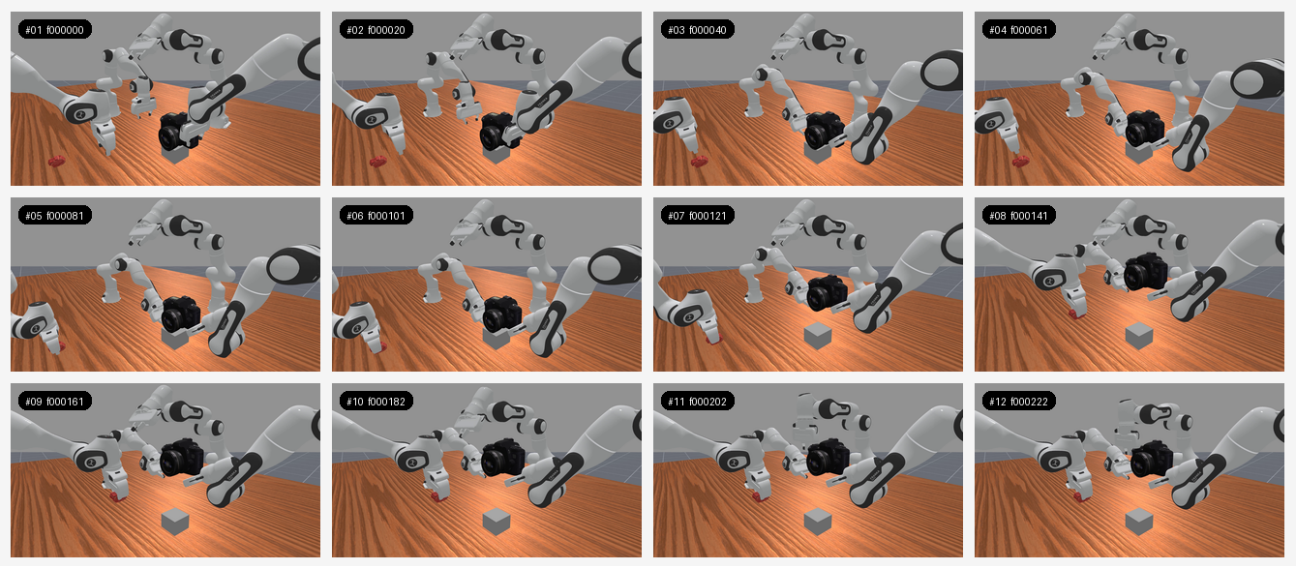

Figure 5. Additional RoboFactory qualitative example (TakePhoto). This successful storyboard complements the Pass-Shoe harder-setting comparison with a distinct multi-agent task.

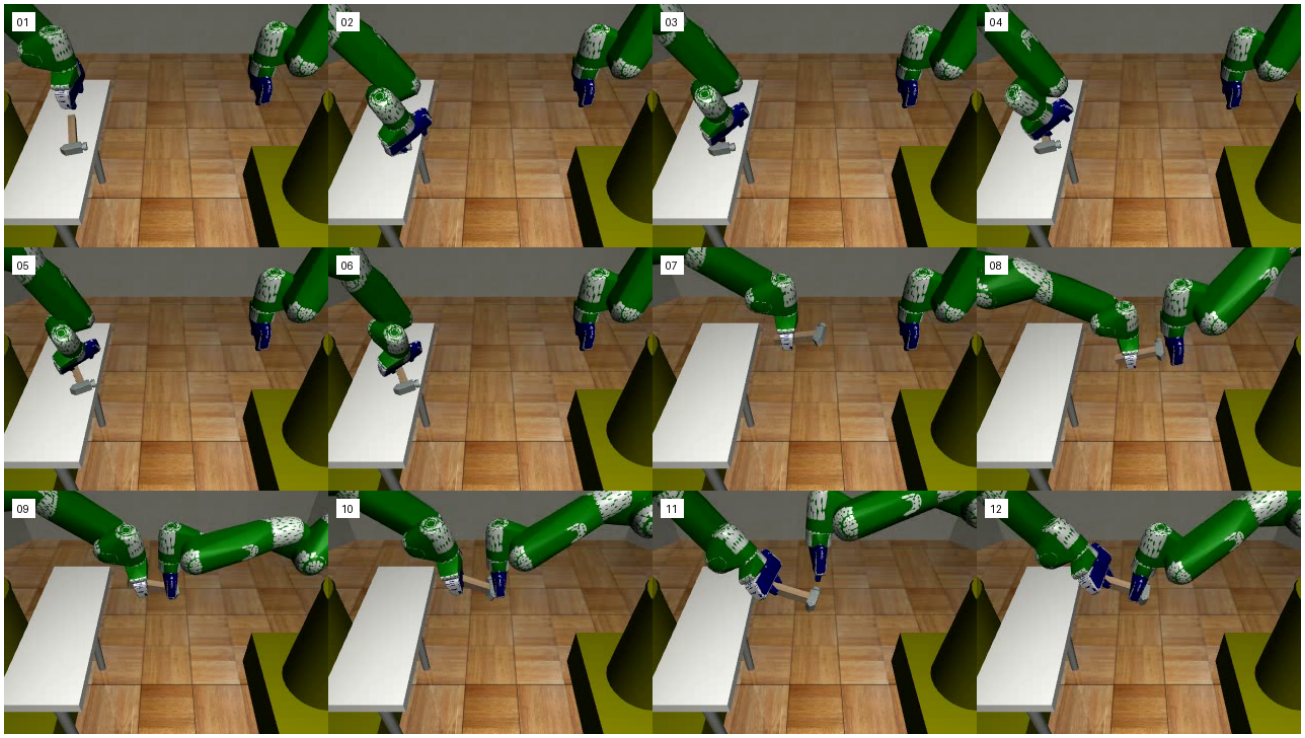

Figure 6. RoboSuite qualitative extension (Two-Arm Handover).

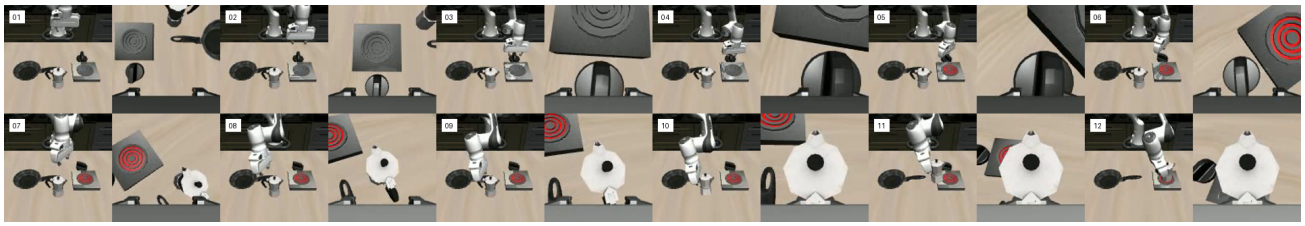

Figure 7. LIBERO Kitchen Scene 3 (Moka Pot task), successful rollout.

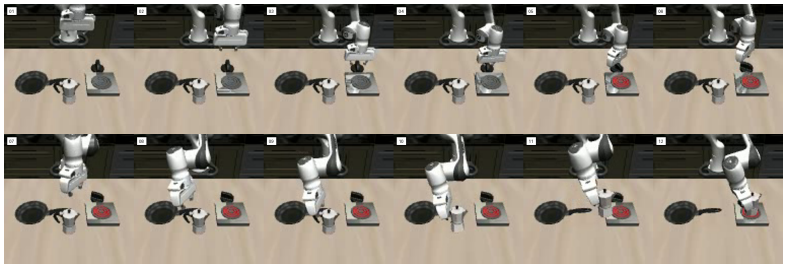

Figure 8. LIBERO Kitchen Scene 3 (Moka Pot task), baseline rollout that fails to complete the task cleanly.

References

Zandieh, A., Daliri, M., Hadian, M., and Mirrokni, V. *TurboQuant: Online Vector Quantization with Near-optimal Distortion Rate*. ICLR 2026.

An, Y., Lu, C., Zhu, K., Yu, T., Zhao, C., Wu, H., Tang, M., and Wang, J. *ReST-KV: Robust KV Cache Eviction with Layer-wise Output Reconstruction and Spatial-Temporal Smoothing*. ICLR 2026.

Liu, G., Li, C., Ning, Z., Lin, J., Yao, Y., Ke, D., Guo, M., and Zhao, J. *FreeKV: Boosting KV Cache Retrieval for Efficient LLM Inference*. ICLR 2026.

Feng, Y., Guo, H., Lv, J., Zhou, S. K., and Xie, X. *Taming the Fragility of KV Cache Eviction in LLM Inference*. ICLR 2026.

Staniszewski, K. and Łańcucki, A. *KV Cache Transform Coding for Compact Storage in LLM Inference*. ICLR 2026.

Luo, X., Wang, Y., Zhu, Q., Zhang, Z., Zhang, X., Yang, Q., and Xu, D. *Turning Trash into Treasure: Accelerating Inference of Large Language Models with Token Recycling*. ACL 2025.

Saranathan, G., Xu, C., Alam, M. P., Kumar, T., Foltin, M., Wong, S. Y., and Bhattacharya, S. *SubLIME: Subset Selection via Rank Correlation Prediction for Data-Efficient LLM Evaluation*. ACL 2025.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., and Lin, J. *Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free*. NeurIPS 2025.

Wang, M., Chu, J., Xie, S., Zang, X., Zhao, Y., and Zhong, W. *Put Teacher in Student's Shoes: Cross-Distillation for Ultra-compact Model Compression Framework*. KDD 2025.

Yu, C., Chen, T., and Gan, Z. *Boost Embodied AI Models with Robust Compression Boundary*. IJCAI 2025.