ProAct

A Benchmark and Multimodal Framework for Structure-Aware Proactive Response

Xiaomeng Zhu[1,2] Fengming Zhu[1] Weijie Zhou[2] Ye Tian[2] Zhenlin Hu[3] Yufei Huang[2]

Yuchun Guo[2] Xinyu Wu[4] Zhengyou Zhang[2] Fangzhen Lin[1] ✉ Xuantang Xiong[2] ✉

[1] Dept. of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong SAR, China

[2] Tencent, Shenzhen, China

[3] Futian Laboratory, Shenzhen, China

[4] Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China

From passive response to proactive response

The robot should pursue the task objective, not merely mirror the human.

Passive agent

Merely follows explicit instructions — mirroring the human's next step.

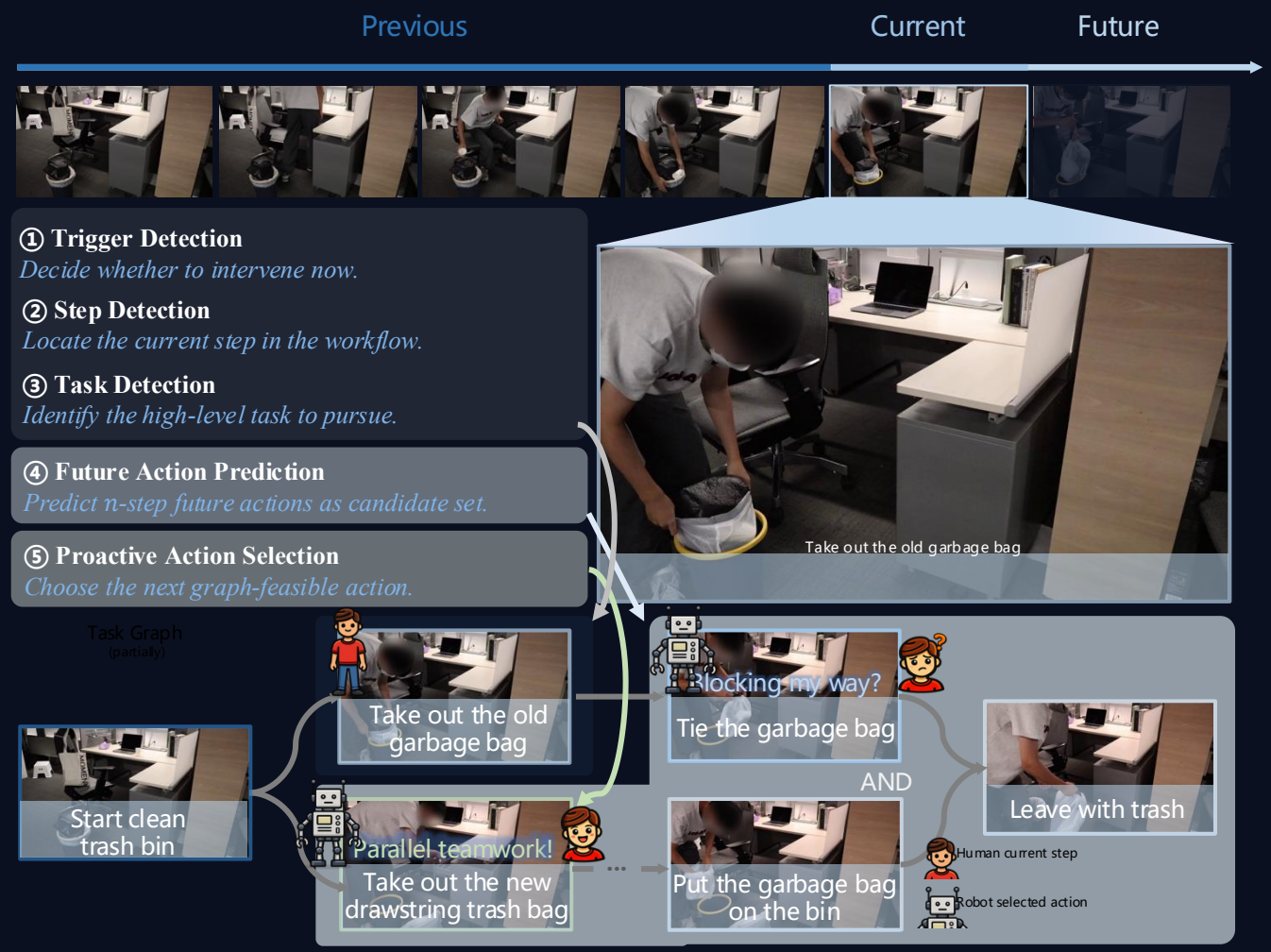


Figure 1. Overview of proactive response — ProAct-75 supports five vision-based tasks with step-level annotations, hierarchical labels, and task graphs.

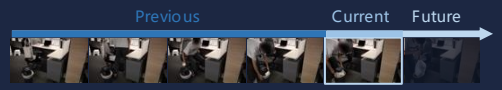
Proactive agent

Aligns with higher-level objectives — keeps watching the scene to decide when and how to act.

Proactive response, formally

At each timestep, the agent reads the scene and the task graph, then outputs five decisions.

INPUT X_t



Video frame
current observation

Task graph
structural prior

OUTPUT Y_t

Trigger	$y_t^{trig} \in \{0,1\}$	act now, or wait?
Task	$y_t^{task} \in \mathcal{T}$	which task category
Step	$y_t^{step} \in V$	the current step
Future actions	$\hat{a}_{t+1:t+n} \in V$	predicted next steps
Proactive action	$a\star_{t+1} \in V_e \cup \{WAIT\}$	the robot's next move

These predictions form a hierarchy — trigger gates task, task scopes step, step grounds the next action.

Why existing benchmarks fall short

Recognition is well studied; structure-aware intervention is not.

What current video benchmarks measure

Action recognition

What action is happening right now

Temporal localization

When does each action start and end

Anticipation

What is the single most likely next action

What proactive response also needs

Should I act now?

Trigger detection — judging the right moment to intervene

Which action stays valid?

Respecting task structure and step dependencies

Can I run in parallel?

Acting on an independent thread, not mirroring the human

No existing benchmark couples perception, anticipation and structured decision-making — ProAct-75 does.

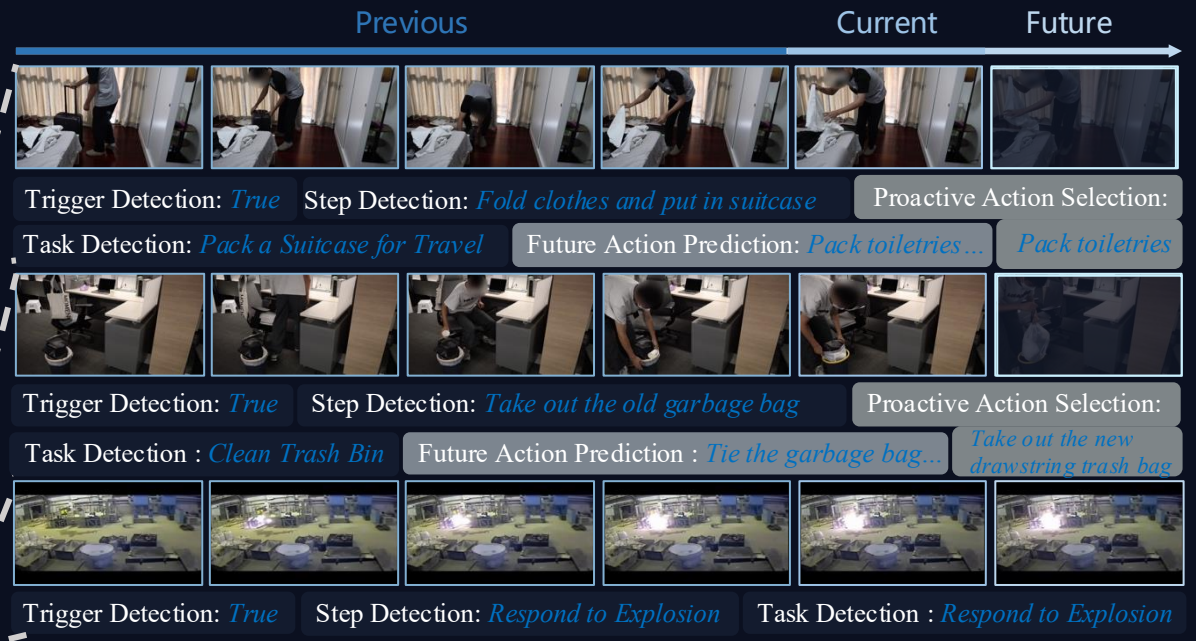
ProAct-75 at a glance

A unified benchmark across assistance, maintenance, and safety monitoring.

Assistance Scenario		
Source	Videos (All / Best)	Annotations (All / Best)
COIN	465 / 465	2,600 / 2,600
Ego-Exo4D	3,446 / 856	85,010 / 19,923
Ours	375 / 125	2,239 / 751

Maintenance Scenario		
Source	(All / Best)	(All / Best)
COIN	93 / 93	552 / 552
Ours	407 / 115	2,381 / 686

Safety Monitoring Scenario		
Source	(All / Best)	(All / Best)
UCF-CRIME	915 / 915	1,053 / 1,053

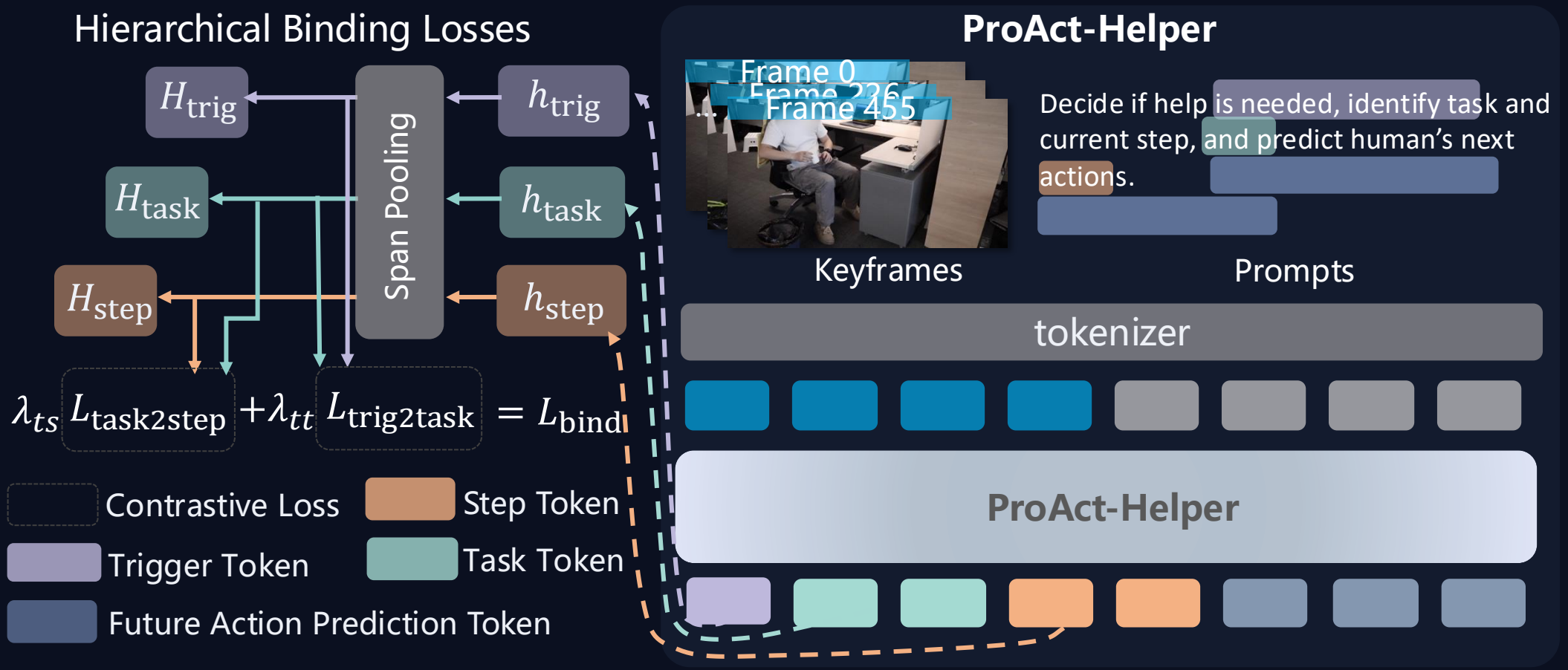


From raw videos to task graphs

A three-stage annotation pipeline turns demonstrations into executable structure.

The ProAct-Helper framework

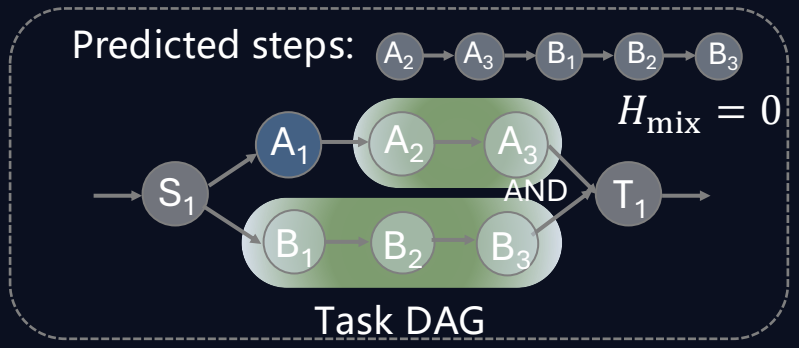
MLLM perception grounds the scene; the task graph constrains planning.



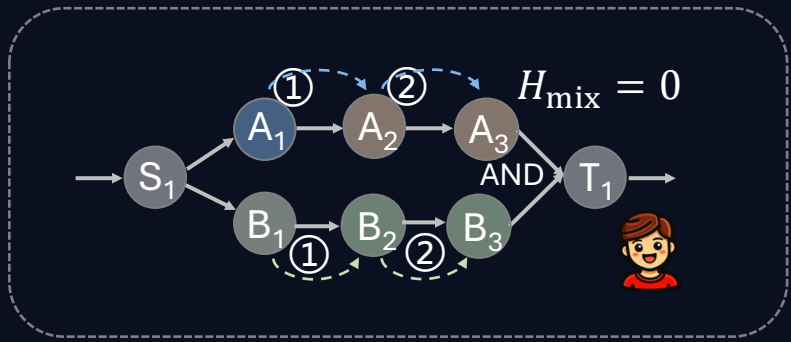
Task graph formulation

A task is a DAG of executable actions and structural (AND / OR) nodes.

Proactive Action Selection



Entropy-driven heuristic search



Human Step Robot Step Candidates

ENTROPY OBJECTIVE

$$a^{\star}_{t+1} = \arg\min_{a} H_{\text{mix}}(H_t, R_t, a)$$

Among graph-feasible actions, pick the one that least mixes the human and robot threads.

Task graph = DAG

Each task is a Directed Acyclic Graph. Executable nodes (V_e) are concrete actions; structural nodes (V_n) carry AND / OR logic.

AND / OR semantics

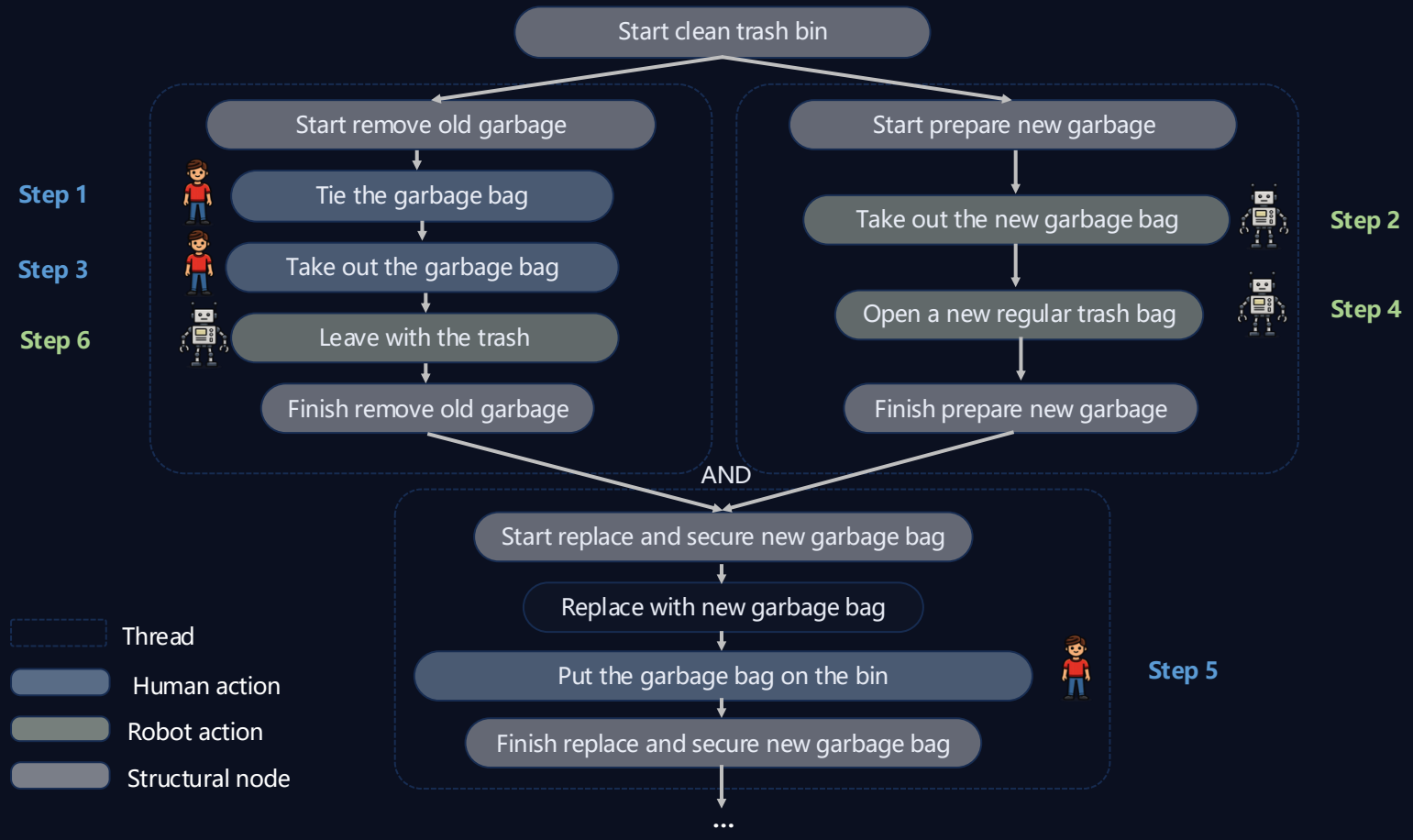
AND node fires only after all predecessors finish; OR node fires once any one of them does.

Parallel threads

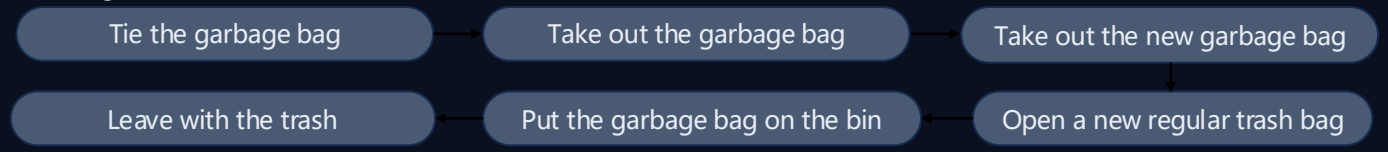
Branches independent until they merge — letting the robot act on its own thread alongside the human.

Task graph: a worked example

A task is a DAG of executable actions and structural (AND / OR) nodes.



Human ground-truth labeled trace:



Main results

ProAct-Helper improves both perception and proactive collaboration.

+6.21%

Trigger detection mF1

▲ over strong closed-source MLLMs

0.25

fewer steps per online decision

▲ more efficient one-step planning

+15.58%

parallel action rate

▲ genuine human–robot parallelism

Better on both axes that matter — perceiving the right moment, and collaborating efficiently.

ProAct: A Benchmark and Multimodal Framework for Structure-Aware Proactive Response

Thank you for watching!

Xiaomeng Zhu[1,2] Fengming Zhu[1] Weijie Zhou[2] Ye Tian[2] Zhenlin Hu[3] Yufei Huang[2]
Yuchun Guo[2] Xinyu Wu[4] Zhengyou Zhang[2] Fangzhen Lin[1] ✉ Xuantang Xiong[2] ✉

[1] Dept. of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong SAR, China

[2] Tencent, Shenzhen, China

[3] Futian Laboratory, Shenzhen, China

[4] Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China