NEURIPS 2026

3A-VLA

Abstraction-Aligned Action Learning

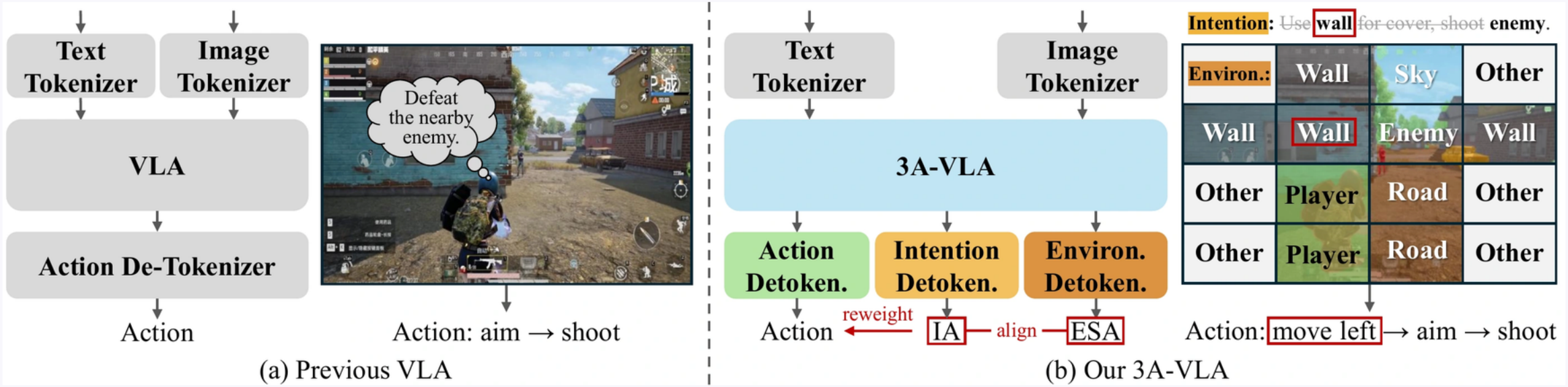
for Vision-Language Agents in 3D Game Worlds

Zheyuan Zhou Liang Du Zixun Sun†
Xiaoyu Zhou Ruimin Ye Qihao Chen Yinda Chen Lemiao Qiu

Tencent IEG x Zhejiang University

† Corresponding author

Action learning with aligned abstractions



During training, the policy learns to connect intent with the scene.

One instruction. The whole scene.

FULL INSTRUCTION

Shoot the nearby enemy.

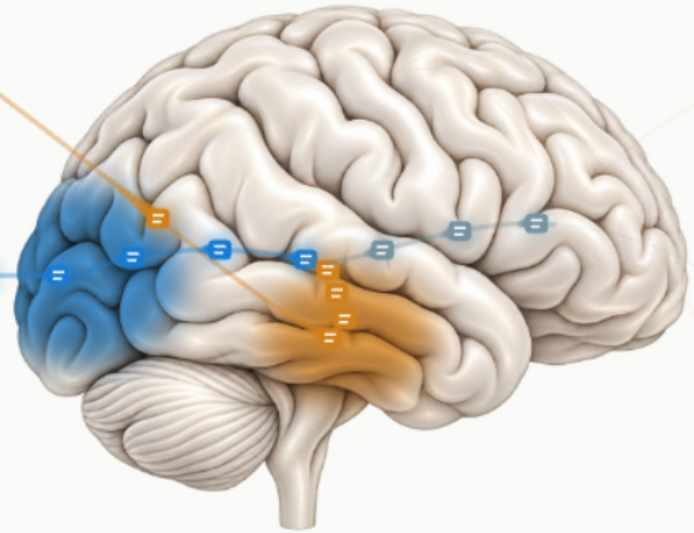
WHOLE IMAGE



Original inputs

MOTOR

GLOBAL WORKSPACE



3A-VLA

One shared model

IA / INTENTION

enemy | wall

ESA / SCENE SEMANTICS



person ↔ enemy cover ↔ wall

ACTION

AIM | SHOOT

IA + ESA: auxiliary training targets

Conceptual illustration

A goal becomes a compact intention.

FULL INSTRUCTION

Shoot the nearby enemy.

WHOLE IMAGE



Original inputs

MOTOR

GLOBAL WORKSPACE



IA / INTENTION

enemy

wall

ESA / SCENE SEMANTICS



person ↔ enemy cover ↔ wall

ACTION

AIM

SHOOT

IA + ESA: auxiliary training targets

3A-VLA

One shared model

Conceptual illustration

Pixels become useful scene structure.

FULL INSTRUCTION

Shoot the nearby enemy.

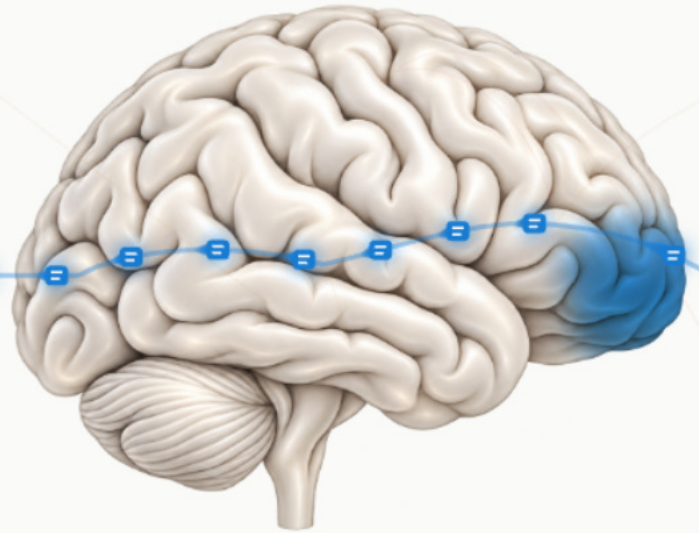
WHOLE IMAGE



Original inputs

MOTOR

GLOBAL WORKSPACE



3A-VLA

One shared model

IA / INTENTION

enemy | wall

ESA / SCENE SEMANTICS



person ↔ enemy cover ↔ wall

ACTION

AIM | SHOOT

IA + ESA: auxiliary training targets

Conceptual illustration

Bring intention and environment together.

FULL INSTRUCTION

Shoot the nearby enemy.

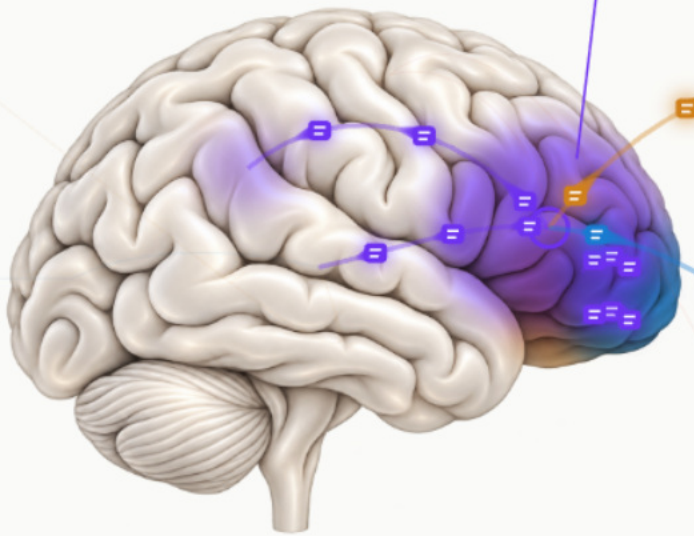
WHOLE IMAGE



Original inputs

MOTOR

GLOBAL WORKSPACE



IA / INTENTION

enemy wall

ESA / SCENE SEMANTICS



person ↔ enemy cover ↔ wall

ACTION

AIM SHOOT

IA + ESA: auxiliary training targets

3A-VLA

One shared model

Conceptual illustration

Use the learned policy to act.

FULL INSTRUCTION

Shoot the nearby enemy.

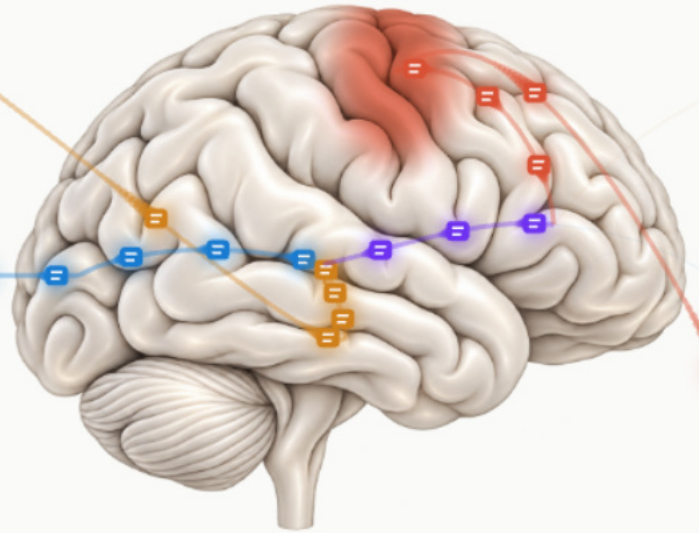
WHOLE IMAGE



Original inputs

MOTOR

GLOBAL WORKSPACE



IA / INTENTION

enemy wall

ESA / SCENE SEMANTICS



person ↔ enemy cover ↔ wall

ACTION

AIM SHOOT

3A-VLA

One shared model

Action-only inference

Conceptual illustration

MAIN RESULTS / TABLE 1

Stronger across three kinds of control

Minecraft MCU benchmark · Mini suite

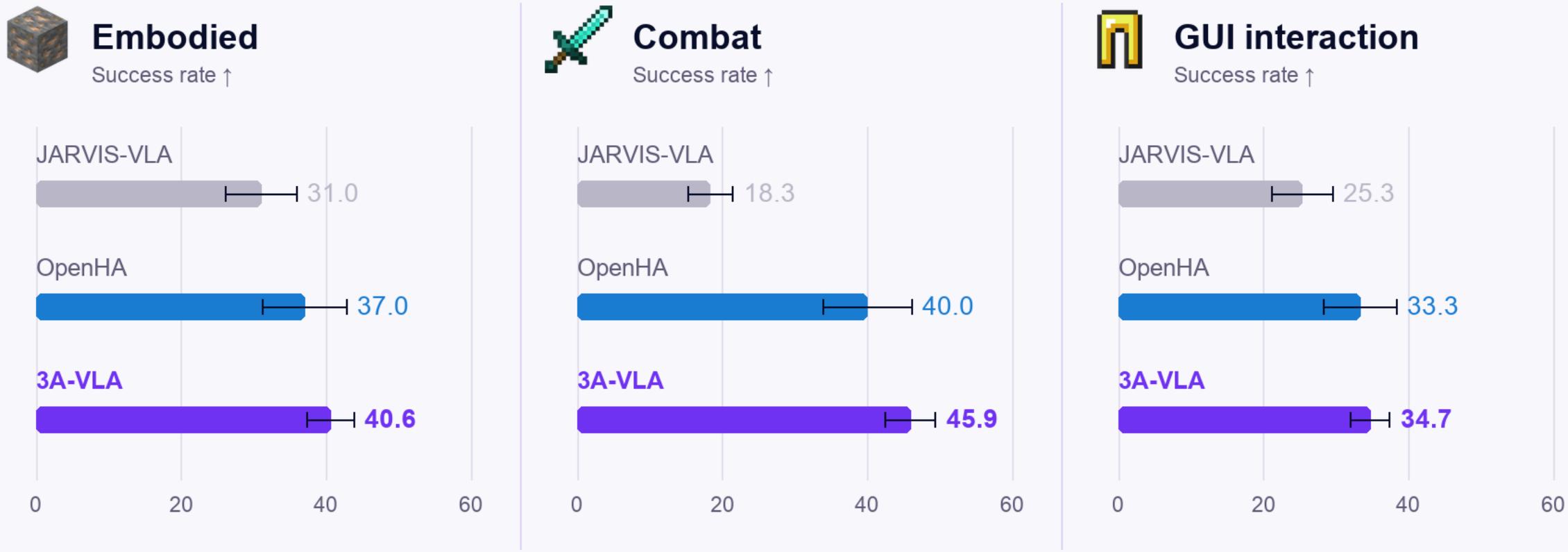


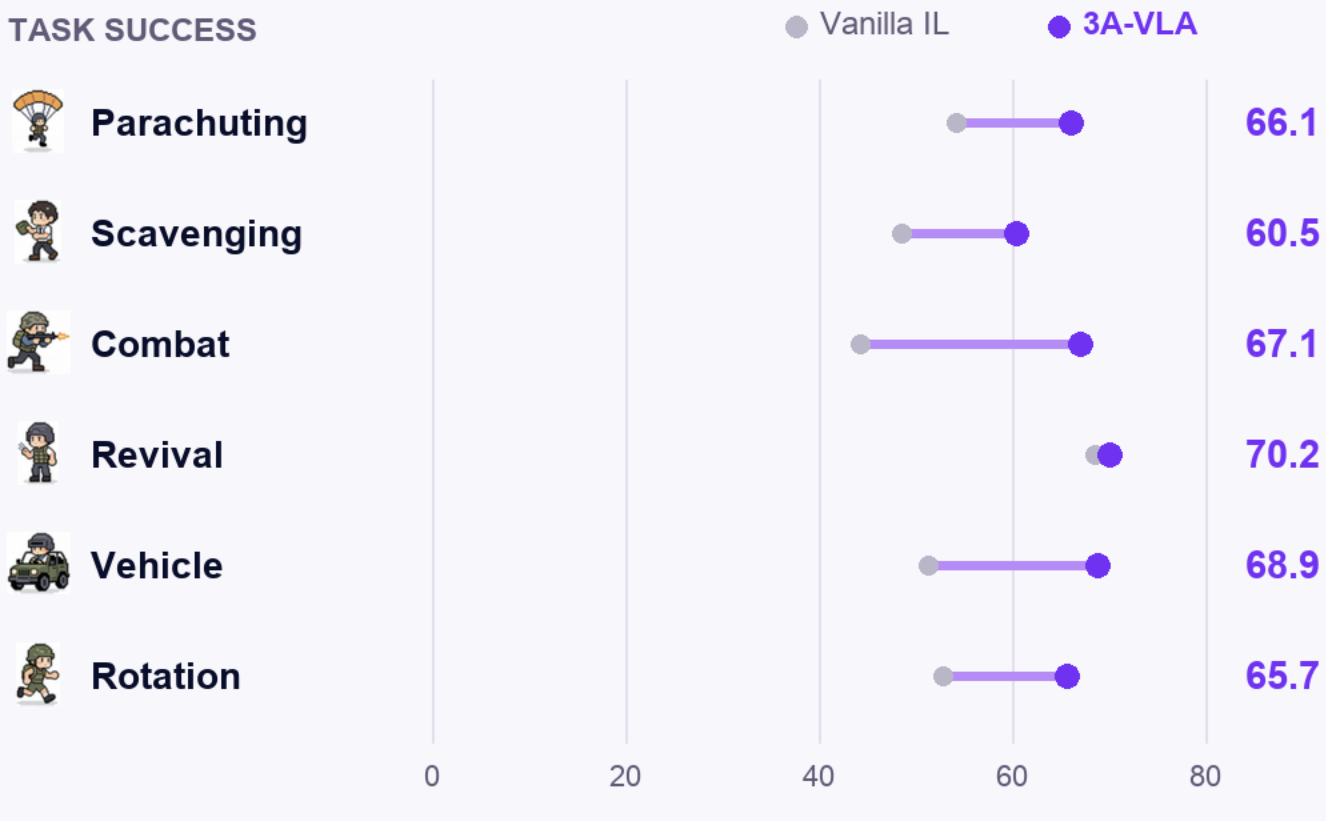
Table 1 · Selected 7B VLA baselines · Success rate (%) · Whiskers show the reported uncertainty

MAIN RESULTS / TABLE 2

Best across all six Game for Peace tasks

Highest success rate among the methods compared in the paper

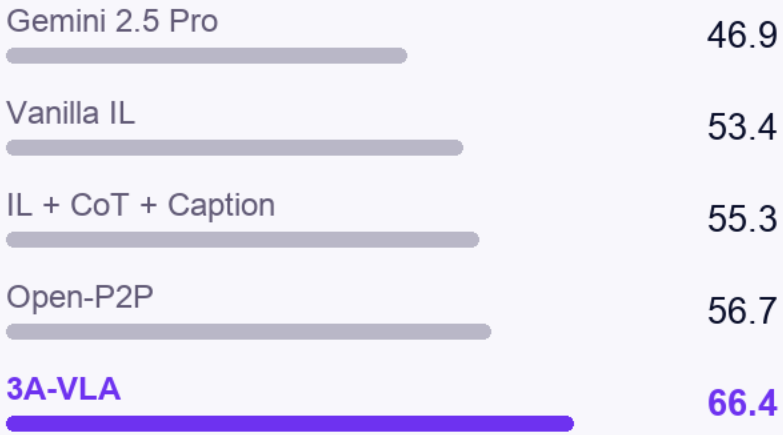
TASK SUCCESS



SIX-TASK AVERAGE

66.4%

+13.0 pp
over Vanilla IL



Same 1.1 s/step as Vanilla IL

Table 2 · Left: Vanilla IL vs. 3A-VLA · Right: selected baselines · No inference-time pruning

COMPONENT ANALYSIS / IA

Give the policy a compact, reusable intent

A concrete supervision example, followed by the controlled Minecraft comparison

ORIGINAL INSTRUCTION

Find a way to quickly enter the safe zone to escape the hazard zone.

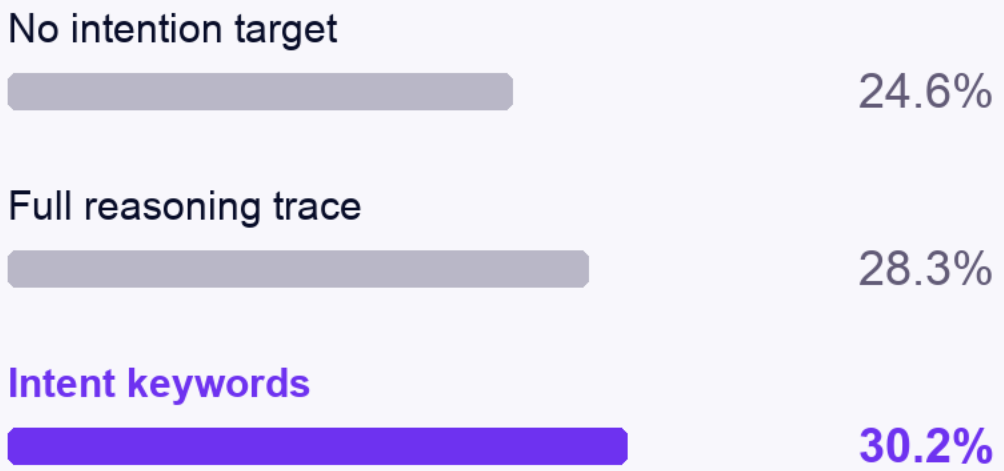
↓

INTENTION TARGET

waypoint **safezone**

Goal and strategy in a few concepts

CHANGE THE TARGET. KEEP THE SETUP.



Compact keywords outperform full reasoning.

Instruction example from the paper · Table 3: IA-only comparison, with ESA and AAR disabled

MINECRAFT EXPERIMENTS / 01B

Latent ESA gives the highest success rate

Minecraft success rate / ESA-only ablation

VISUAL SUPERVISION TARGET	SUCCESS RATE
Image caption	27.0%
Full-image prediction	26.8%
Full-image semantic segmentation	27.4%
Latent scene semantics (ESA)	30.6%

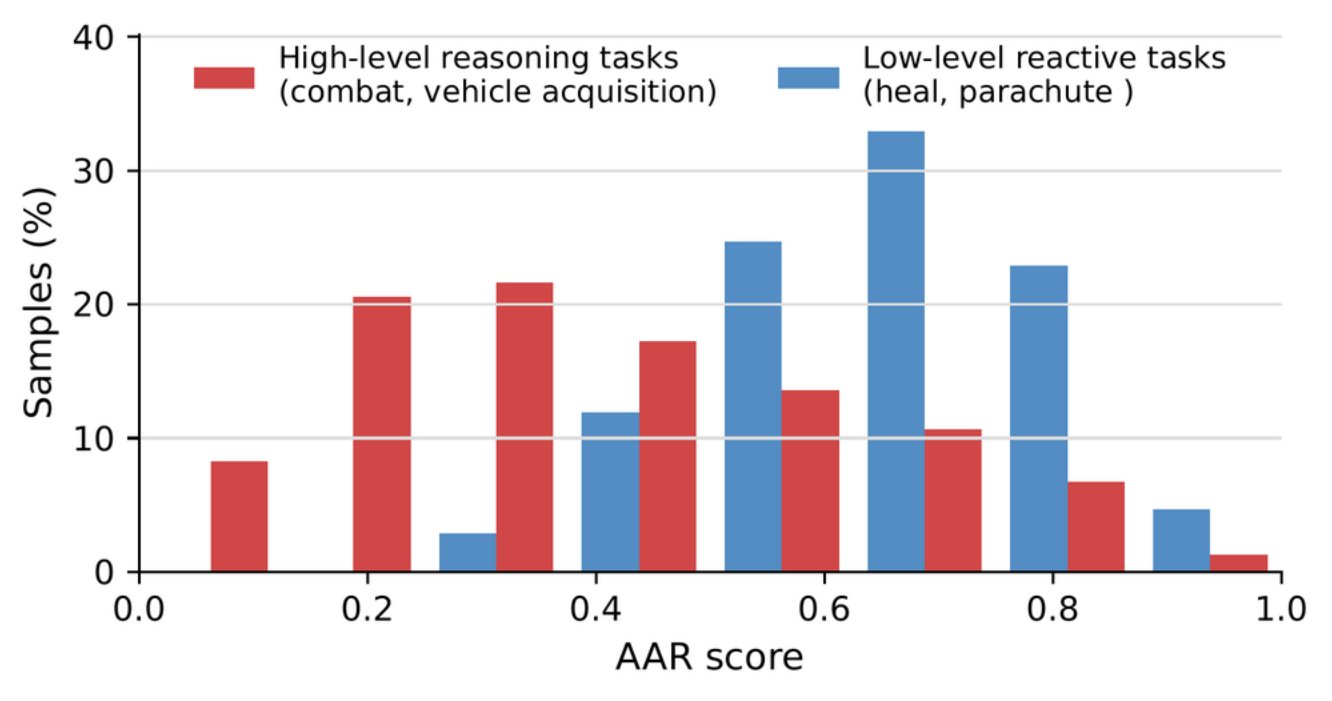
Camera-ready main ablation · IA and AAR are disabled for all four rows

COMPONENT ANALYSIS / AAR

Learn which demonstrations to emphasize

Intention–scene alignment determines how strongly an example teaches the action policy

OBSERVED ALIGNMENT ACROSS TASK TYPES



WHAT THE RELATION TELLS US

Stronger alignment

The scene more directly supports the intended low-level action.

→ **Greater action-loss weight**

Weaker alignment

The same observation may admit several plausible actions.

→ **Lower action-loss weight**

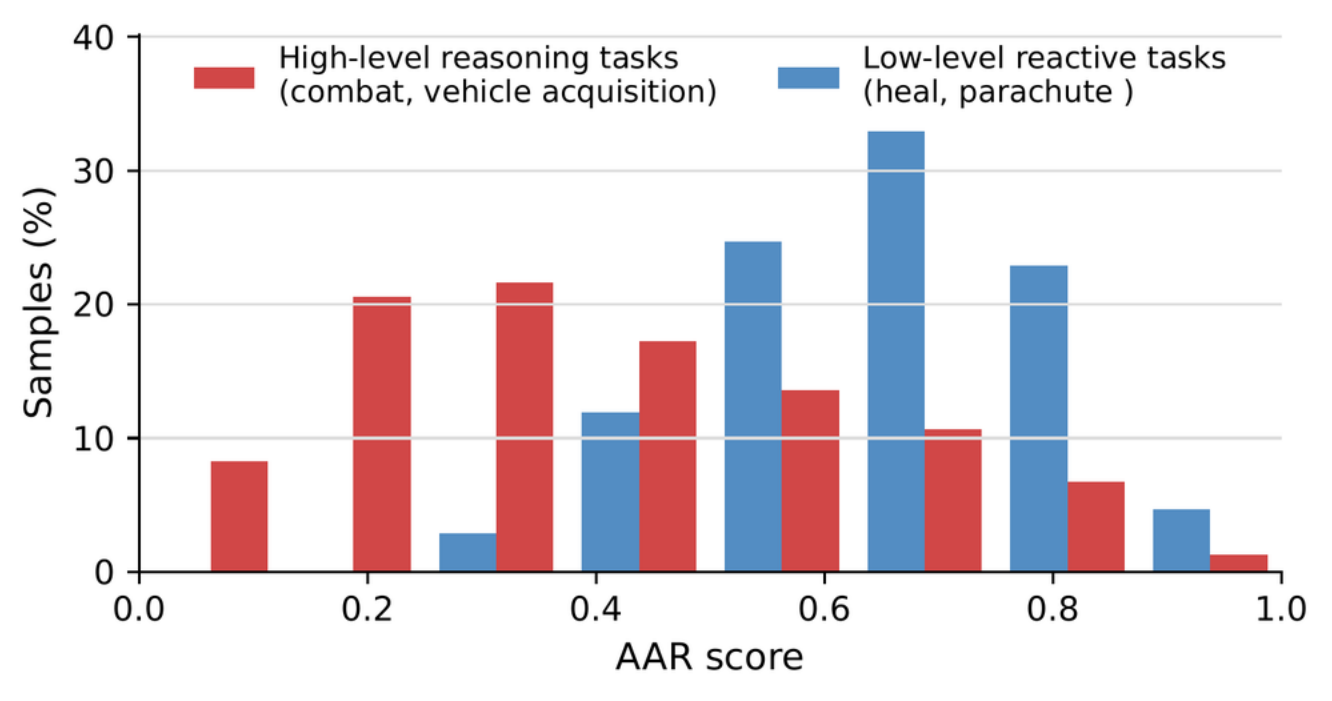
Original AAR score distributions and interpretation from the appendix analysis

COMPONENT ANALYSIS / AAR

Learn which demonstrations to emphasize

Intention–scene alignment determines how strongly an example teaches the action policy

OBSERVED ALIGNMENT ACROSS TASK TYPES



SAME TARGETS. DIFFERENT WEIGHTING.

Minecraft success rate (%)

Without AAR	36.5
Fixed gate 0.33	36.9
Fixed gate 0.67	37.0
Scalar cosine	37.1
Scalar Hadamard	37.6
Learned AAR	**40.4**

Learned weights improve action learning.

Left: original AAR score distributions, Appendix analysis · Right: weighting controls, Table 4

Focused attention opens the door to speed

Task-relevant regions carry a larger share of the attention

Vanilla imitation



3A-VLA



VISUAL TOKENS

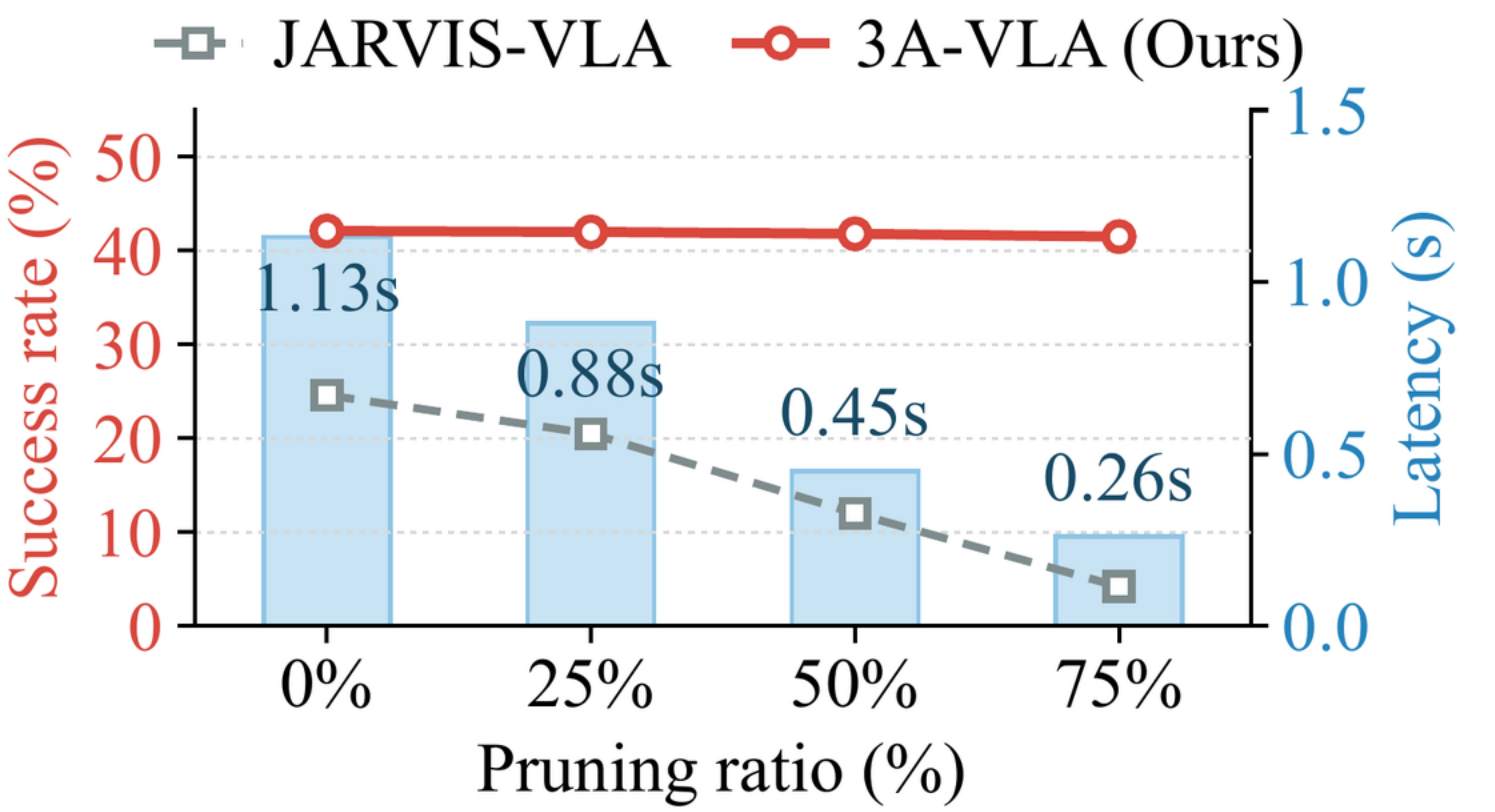
Keep the most connected

Attention examples: original Figure 3 · Token selection below is a schematic, not a measured pruning mask

DEPLOYMENT ANALYSIS / PRUNING

Remove tokens while retaining task success

A separate Minecraft study with enhanced IA + ESA



75% OF TOKENS REMOVED

Latency per step

1.1 s → 0.3 s

Success rate

42.1% → 41.5%

Faster inference.

A small success-rate tradeoff.

Original pruning figure and comparison table · Latency cards use the table's rounded values

GAME FOR PEACE / CONTINUOUS ROLLOUT

Parachuting and scavenging

One continuous rollout, from the air to collecting a weapon



01 Parachute



02 Land



03 Collect a weapon

Watch the demonstrations on the project page

3a-vla.github.io

CROSS-GAME / QUALITATIVE SHOWCASE

Driving

Vehicle control across different game worlds

Game for Peace



GTA V



Watch the demonstrations on the project page

3a-vla.github.io

CROSS-GAME / QUALITATIVE SHOWCASE

Combat

Combat behavior across different game worlds

Game for Peace



Valorant



Watch the demonstrations on the project page

3a-vla.github.io

CROSS-GAME / QUALITATIVE SHOWCASE

Traversal

Movement across different game worlds

Game for Peace



Elden Ring



Watch the demonstrations on the project page

3a-vla.github.io

3A-VLA

Abstraction-Aligned Action Learning

for Vision-Language Agents in 3D Game Worlds



3a-vla.github.io

Model · Code · Dataset