# CombatVLA: An Efficient Vision-Language-Action Model for Combat Tasks in 3D Action Role-Playing Games

Supplementary Material

## 1. Overview



## 2. Limitations

We must also candidly acknowledge some limitations in our research, specifically: 1) Task Definitions: As VLM- and VLA-based agents are still evolving, the current task definitions are somewhat simplistic. 2) Game Scenarios: Our research has only been tested within the BMW and SSDT game and has not yet been extended to other scenarios. 3) Model Capabilities: As shown in the benchmark evaluation section, there is still room for improvement in existing VLMs and VLAs.

## 3. More Details

### 3.1. Details of Data Annotation

The game annotation team comprises six individuals, each of whom has completed all levels of the game. Over a two-week period, their gameplay data was recorded using our action tracker. After filtering out abnormal samples with insufficient action density, we cleaned 200 hours of recordings, including video, mouse, and keyboard inputs.

The data annotation team consists of ten members, each with at least a bachelor's degree and gaming experience. They are responsible for annotating the benchmark data and the formatted AoT data generated by GPT-4o. All QA pairs are annotated by the this team and cross-validated to ensure high quality. This validation process ensures that only data passing all quality checks is retained.

Ultimately, we compiled 914 data fragments for our CUBench, 25,000 images, and 5,000 high-quality AoTs.

### 3.2. Details of Prompts

The prompts of QA pair generation are as follows,

**Prompts of Benchmark Collection**

**—— Gathering ——**

**Gathering enemy health**

Select the best answer to the following single-choice question based on the game-screenshot image. Respond with only the letter (Yes or No) of the correct option. Is the enemy's health high in the game? Yes/No. The best answer is:

**Gathering own health**

Select the best answer to the following single-choice question based on the game-screenshot image. Respond with only the letter (Yes or No) of the correct option.Is the health of the game character you control high in the game? Yes/No. The best answer is:

**Gathering own abnormal status**

Select the best answer to the following single-choice question based on the game-screenshot image. Respond with only the letter (Yes or No) of the correct option. Is the game character in an abnormal state? (Such as being on fire) Yes/No. The best answer is:

**—— Comprehension ——**

**Understanding action intention**

Select the best answer to the following single-choice question based on the game-screenshot image. Respond with only the letter (Yes or No) of the correct option.Carefully observe the enemy's movements. Will the enemy attack next or is it attacking now? Yes/No. The best answer is:

**Understanding current state**

Select the best answer to the following single-choice question based on the game-screenshot image. Respond with only the letter (Yes or No) of the correct option.Is the enemy in a stunned state? (When the enemy is in a stunned state, they cannot attack for a period of time and can only be attacked. For example, the enemy is knocked down or immobilized by the spell.) Yes/No. The best answer is:

**—— Reasoning ——**

Q: Select the best answer to the following single-choice question based on the game-screenshot image. Respond with only the letter (A, B, or C) of the correct option.Carefully observe the enemy's actions. As the game character, please reason which of the following actions is most suitable for your next move (ensure your health is prioritized while depleting the enemy's health). A. Restore health of the game character. B. Dodge to avoid enemy attacks and prevent damage. C. Attack the enemy. The best answer is:


Figure 1. The visualization of 13 defined tasks.

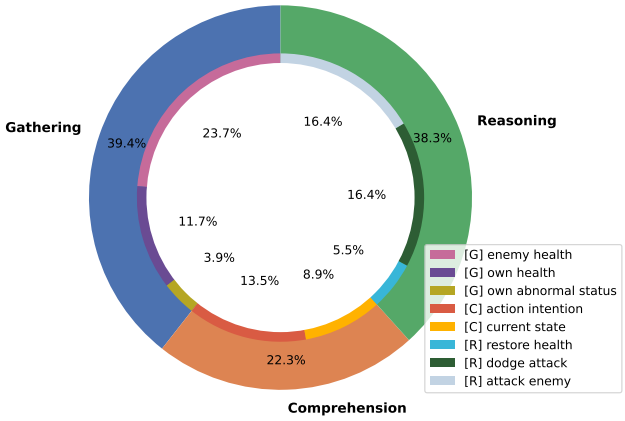


Figure 2. Distribution of 3 tasks (i.e., gathering, understanding, reasoning) and their 8 subtasks.

### 3.3. Details of CUBench Benchmark

To thoroughly assess the combat IQ of our CombatVLA and all baselines, we developed CUBench. As illustrated in Fig.2, this benchmark is composed of three types of tasks: 39.4% gathering, 22.3% understanding, and 38.3% reasoning. Each of these main tasks is further divided into 8 subtasks. Tab.1 presents a detailed breakdown.

| Task Category | Volume |
|---|---|
| **Gathering** | **360** |
| Gathering enemy health | 217 |
| Gathering own health | 107 |
| Gathering own abnormal status | 36 |
| **Comprehension** | **204** |
| Understanding action intention | 123 |
| Understanding current state | 81 |
| **Reasoning** | **350** |
| Option A: restore health | 50 |
| Option B: dodge attack | 150 |
| Option C: attack enemy | 150 |

Table 1. Benchmark statistics of CUBench.

### 3.4. Additional Quantitative Results

For comparison with traditional RL-based methods, we used the state-of-the-art RL-based method, AI-Wukong, as the baseline in task-level practical tests. It takes visual images as input and outputs fixed and predefined action combinations, training ResNet with DQN/PPO. The settings align with CombatVLA (training on Task 9-10, zero-shot testing


Figure 3. Additional qualitative visualization of actions and corresponding frame sequences.

on others).

The results in Tab. 2 show that RL-based method, despite having low latency, only succeeds in the training domain (Task 9) and easy tests (Task 1), failing completely on other tasks. Moreover, its online training requires hundreds of character deaths, takes 42 hours to train. In stark contrast, CombatVLA trains in **just 4 hours**. This limitation arises because RL-based method depend on manually labeled specific tasks, which hampers their ability to generalize. In contrast, CombatVLA genuinely comprehends 3D combat rather than merely executing a visual classification task.

### 3.5. Additional Qualitative Visualization

Fig. 3 illustrates the visualization highlights of additional combat tasks. In the first row, CombatVLA moves the game character away from the enemy before restoring health to ensure its own safety. The second and third rows show that CombatVLA charges forward to perform a series of consecutive attacks immediately after immobilizing the enemy. In the fourth row, the enemy's attacks can only be dodged by moving left or right or rolling, so CombatVLA first moves left and then rolls to evade. This indicates that through progressive learning, it has learned the enemy's attack patterns in task 9. In the fourth row, CombatVLA is able to precisely block an enemy's attack, demonstrating strong generalization capability even in zero-shot tests of different games. These cases prove that CombatVLA can make the right decisions at the right time.

### 3.6. Additional Ablation Study

We have included ablation experiments on task-level practical tests, where we selected a representative task from each difficulty level to evaluate the success rate and reported the inference time. As shown in Tab.3, our setting achieves optimal performance.

### 3.7. Task Defination

As shown in Fig. 1, which corresponds to Tab.1 in the main text, is a visual representation of the defined tasks. The first two rows are tasks from BMW, and the last row features tasks from SSDT. The enemies in these tasks vary in appearance, attack patterns, health, and skills, which will thoroughly test the robustness of VLAs in combat tasks.

Table 2. Quantitative comparison with the RL-based method in task-level practical tests.

| Model | Task1(Easy) | Task6(Middle) | Task8(Hard) | Task9(Very Hard) | Task13(Hard) | Avg. | Training Time(h)↓ | Latency(s)↓ |
|---|---|---|---|---|---|---|---|---|
| AI-Wukong(RL) | 20% | 0% | 0% | 10% | 0% | 6% | 42.3 | **0.23** |
| CombatVLA | **100%** | **80%** | **60%** | **10%** | **60%** | **62%** | **4.1** | 1.85 |

Table 3. Ablation study on task-level practical tests.

| Setting | Task1(Easy) | Task6(Middle) | Task8(Hard) | Task9(Very Hard) | Task13(Hard) | Avg. | Time(s)↓ |
|---|---|---|---|---|---|---|---|
| Training Stage1 | 60% | 50% | 20% | 0% | 30% | 32% | 3.73 |
| Training Stage2 | 100% | 70% | 60% | 10% | 40% | 56% | 3.73 |
| w/o $\mathcal{L}_{con}$ | 100% | 60% | 50% | 0% | 20% | 46% | 1.85 |
| w/o $\mathcal{L}_{align}$ | 90% | 70% | 50% | 0% | 30% | 48% | 1.85 |
| CombatVLA | **100%** | **80%** | **60%** | **10%** | **60%** | **62%** | **1.85** |

## 4. Ethical Consideration

Our approach can automate the playing of ARPGs, which may lead to cheating and false advertising in games, potentially causing negative impacts on society. Therefore, it is crucial to develop methods capable of distinguishing between authentic and fraudulent content. We strongly condemn unauthorized and malicious use of this technology and emphasize the importance of considering ethical issues when utilizing our approach.

## 5. Demo Video

We have provided a detailed demo video to demonstrate the effectiveness of our CombatVLA. The first video is a full demonstration of CombatVLA completing tasks 1 through 13. For smoother viewing, we have edited out the game pauses. The second video is a comparison of inference speeds between CombatVLA and VARP. We have kept the game pauses in this video, with the pause duration representing the inference time taken by the two methods. The video demonstrates that our method is significantly faster than VARP. Please refer to the website https://combatvla.github.io/.