Formal Specification: Two-Player 3D Grid-World "Weighted-Target" Problem

Below is a fully-formal specification of the two-player 3D Grid-World "weighted-target" problem, **under the assumption that each agent**:

1. **Knows its own initial position** $p_{\text{init}}^{(i)}$.
2. **Knows the opponent's initial position** $p_{\text{init}}^{(j)}$.
3. **Observes exactly the opponent's last action** $\Delta_{t-1}^{(j)}$ at each turn.
4. **Does not observe the opponent's current position** $p_t^{(j)}$. Instead, it must *estimate* $p_t^{(j)}$ from its knowledge of initial positions, its own actions, and the sequence of observed opponent-actions so far. All other aspects (collision, weighted targets, turn order, etc.) remain as in the standard 3D Grid-World. We cast this as a **two-player partially-observable turn-based Markov game**.

Notation

- $X, Y, Z \in \mathbb{N}$: dimensions of the 3D grid.
- $W = \{\, (x, y, z) \in \mathbb{Z}^3 \mid 0 \le x < X,\ 0 \le y < Y,\ 0 \le z < Z\}$.
- Player indices: $i \in \{1, 2\}$, and $j = 3 - i$ denotes "the other player."
- Initial (hidden) positions:

$$p_{\text{init}}^{(i)} \ \in \ W, \quad p_{\text{init}}^{(j)} \ \in \ W.$$

  Each agent i knows both $p_{\text{init}}^{(i)}$ and $p_{\text{init}}^{(j)}$ from the start.
- Weighted targets:

$$T_{\text{init}} = \{\tau_1, \tau_2, \ldots, \tau_n\} \subset W, \qquad V(\tau_k) \in \mathbb{Z}_{>0} \text{ for each } k.$$

  At time t, the uncollected targets form $T_t \subseteq T_{\text{init}}$.
- Player i's cumulative score at time t: $S_t^{(i)} \in \mathbb{Z}$, initially $S_0^{(i)} = 0$.
- Collision penalty: if a player attempts to move into the other's current cell, that player is "bounced" one cell backward and incurs a -1 penalty.
- Action-vectors:

$$\Delta_t^{(i)} = \left(\Delta x_t^{(i)}, \Delta y_t^{(i)}, \Delta z_t^{(i)}\right) \ \in \ \{-1, 0, 1\}^3,$$

  meaning "one-step attempt" (or $(0, 0, 0)$ for "stay"). —

1. Full-State Space $\mathcal{S}$

A *true* (hidden) state at time t is

$$s_t \ = \ \left(p_t^{(1)},\ p_t^{(2)},\ T_t,\ S_t^{(1)},\ S_t^{(2)}\right),$$

where

1. $p_t^{(i)} \in W$ is the current cell of Player i.
2. $T_t \subseteq T_{\text{init}}$ is the set of targets not yet collected.
3. $S_t^{(i)} \in \mathbb{Z}$ is Player i's score. **Initial state** s_0 is specified by

$$p_0^{(i)} = p_{\text{init}}^{(i)}, \quad T_0 = T_{\text{init}}, \quad S_0^{(i)} = 0 \ (i = 1, 2).$$

—

2. Actions $\mathcal{A}^{(i)}$

At each time-step t, Player i chooses

$$\Delta_t^{(i)} \ = \ (\Delta x_t^{(i)}, \Delta y_t^{(i)}, \Delta z_t^{(i)}) \ \in \ \{-1, 0, 1\}^3.$$

* If $\Delta_t^{(i)} = (0, 0, 0)$, that is "stay in place." * Otherwise, the intended forward-move is

$$\big(x_t^{(i)} + \Delta x_t^{(i)},\ y_t^{(i)} + \Delta y_t^{(i)},\ z_t^{(i)} + \Delta z_t^{(i)}\big),$$

which is then clamped coordinate-wise into $[0..\,X-1] \times [0..\,Y-1] \times [0..\,Z-1]$. Denote

$$\mathcal{A}^{(i)} = \{-1, 0, 1\}^3, \quad i = 1, 2.$$

—

3. Turn-Order and Transition Dynamics

Each full time-step $t = 0, 1, 2, \ldots$ consists of two *ordered sub-steps*:

1. **Sub-step $t.1$ (Player 1 moves)** using $\Delta_t^{(1)}$.
2. **Sub-step $t.2$ (Player 2 moves)** using $\Delta_t^{(2)}$, now against Player 1's updated cell.

Below we write $s_t = \big(p_t^{(1)}, p_t^{(2)}, T_t, S_t^{(1)}, S_t^{(2)}\big)$.

3.1. Sub-step $t.1$: Player 1's Move

1. **Compute forward-clamped position**

$$\widetilde{p}^{(1)} \ = \ \Big(\operatorname{clamp}\big(x_t^{(1)}+\Delta x_t^{(1)},\ 0,\ X-1\big), \operatorname{clamp}\big(y_t^{(1)}+\Delta y_t^{(1)},\ 0,\ Y-1\big), \operatorname{clamp}\big(z_t^{(1)}+\Delta z_t^{(1)},\ 0,\ Z-1\big)\Big).$$

2. **Collision check** (with Player 2 at $p_t^{(2)}$):
   - If $\widetilde{p}^{(1)} \neq p_t^{(2)}$, then **no collision**. Set

$$p_{t+\frac{1}{2}}^{(1)} \ = \ \widetilde{p}^{(1)}, \quad S_{t+\frac{1}{2}}^{(1)} \ = \ S_t^{(1)}.$$

   - If $\widetilde{p}^{(1)} = p_t^{(2)}$, then **collision**:

1. Bounce backwards by $-\Delta_t^{(1)}$ (clamped):

$$\widehat{p}^{(1)} = \Big(\operatorname{clamp}\big(x_t^{(1)}-\Delta x_t^{(1)},\,0,\,X-1\big),\,\operatorname{clamp}\big(y_t^{(1)}-\Delta y_t^{(1)},\,0,\,Y-1\big),\,\operatorname{clamp}\big(z_t^{(1)}-\Delta z_t^{(1)},\,0,\,Z-1\big)$$

2. Then

$$p^{(1)}_{t+\frac{1}{2}} = \widehat{p}^{(1)}, \quad S^{(1)}_{t+\frac{1}{2}} = S_t^{(1)} - 1.$$

Meanwhile, Player 2 does nothing in this sub-step:

$$p^{(2)}_{t+\frac{1}{2}} = p_t^{(2)}, \quad S^{(2)}_{t+\frac{1}{2}} = S_t^{(2)}.$$

3. **Target collection by Player 1** If $p^{(1)}_{t+\frac{1}{2}} \in T_t$, say $p^{(1)}_{t+\frac{1}{2}} = \tau_k$, then Player 1 collects τ_k:

$$S^{(1)}_{t+\frac{1}{2}} \;=\; S^{(1)}_{t+\frac{1}{2}} + V(\tau_k), \quad T_{t+\frac{1}{2}} = T_t \setminus \{\tau_k\}.$$

Otherwise, $T_{t+\frac{1}{2}} = T_t$.

After sub-step $t.1$, we have the *intermediate state*

$$s_{t+\frac{1}{2}} = \Big(p^{(1)}_{t+\frac{1}{2}},\, p^{(2)}_{t+\frac{1}{2}},\, T_{t+\frac{1}{2}},\, S^{(1)}_{t+\frac{1}{2}},\, S^{(2)}_{t+\frac{1}{2}}\Big).$$

—

3.2. Sub-step $t.2$: Player 2's Move

Starting from $s_{t+\frac{1}{2}}$, let

1. **Forward-clamped position**

$$\widetilde{p}^{(2)} = \Big(\operatorname{clamp}\big(x_t^{(2)}+\Delta x_t^{(2)},\,0,\,X-1\big),\,\operatorname{clamp}\big(y_t^{(2)}+\Delta y_t^{(2)},\,0,\,Y-1\big),\,\operatorname{clamp}\big(z_t^{(2)}+\Delta z_t^{(2)},\,0,\,Z-1\big)\Big).$$

2. **Collision check** (against Player 1's updated cell $p^{(1)}_{t+\frac{1}{2}}$):
   - If $\widetilde{p}^{(2)} \neq p^{(1)}_{t+\frac{1}{2}}$, **no collision**:

$$p^{(2)}_{t+1} = \widetilde{p}^{(2)}, \quad S^{(2)}_{t+1} = S^{(2)}_{t+\frac{1}{2}}.$$

   Meanwhile, $p^{(1)}_{t+1} = p^{(1)}_{t+\frac{1}{2}}$ and $S^{(1)}_{t+1} = S^{(1)}_{t+\frac{1}{2}}$.

   - If $\widetilde{p}^{(2)} = p^{(1)}_{t+\frac{1}{2}}$, **collision**:
     1. Bounce backwards:

$$\widehat{p}^{(2)} = \Big(\operatorname{clamp}\big(x_t^{(2)}-\Delta x_t^{(2)},\,0,\,X-1\big),\,\operatorname{clamp}\big(y_t^{(2)}-\Delta y_t^{(2)},\,0,\,Y-1\big),\,\operatorname{clamp}\big(z_t^{(2)}-\Delta z_t^{(2)},\,0,\,Z-1\big)$$

2. Then

$$p^{(2)}_{t+1} = \widehat{p}^{(2)}, \quad S^{(2)}_{t+1} = S^{(2)}_{t+\frac{1}{2}} - 1,$$

while $p^{(1)}_{t+1} = p^{(1)}_{t+\frac{1}{2}}$, $S^{(1)}_{t+1} = S^{(1)}_{t+\frac{1}{2}}$.

3. **Target collection by Player 2** If $p^{(2)}_{t+1} \in T_{t+\frac{1}{2}}$, say $p^{(2)}_{t+1} = \tau_m$, then

$$S^{(2)}_{t+1} = S^{(2)}_{t+1} + V(\tau_m), \quad T_{t+1} = T_{t+\frac{1}{2}} \setminus \{\tau_m\}.$$

Otherwise, $T_{t+1} = T_{t+\frac{1}{2}}$.

At the end of sub-step $t.2$, we arrive at the new global state

$$s_{t+1} = \left(p^{(1)}_{t+1},\, p^{(2)}_{t+1},\, T_{t+1},\, S^{(1)}_{t+1},\, S^{(2)}_{t+1}\right).$$

Because all updates are deterministic given $(s_t, \Delta^{(1)}_t, \Delta^{(2)}_t)$, the transition kernel $P(s_{t+1} \mid s_t, \Delta^{(1)}_t, \Delta^{(2)}_t)$ is a point-mass on this unique s_{t+1}. —

4. Reward Functions

At the end of full time-step t, Player i receives reward

$$r^{(i)}_{t+1} = S^{(i)}_{(\text{after } i \text{ moved})} \; - \; S^{(i)}_{(\text{just before } i \text{ moved})} \; \in \; \{-1,\, 0,\, +V(\tau)\}.$$

* **If Player i collides** on its sub-step, then $S^{(i)}$ decreased by 1, so $r^{(i)}_{t+1} = -1$. * **If Player i collects** a target τ of value $V(\tau)$, then $r^{(i)}_{t+1} = +V(\tau)$. * **Otherwise**, $r^{(i)}_{t+1} = 0$. Specifically:

1. $r^{(1)}_{t+1} = S^{(1)}_{t+\frac{1}{2}} - S^{(1)}_t$.
2.

$$r^{(2)}_{t+1} = S^{(2)}_{t+1} - S^{(2)}_{t+\frac{1}{2}}.$$

5. Observation Spaces $\mathcal{O}^{(i)}$

Because each agent **does know both initial positions** but **does not see the opponent's current position**, we define:

- At the start of time-step t, Player i has just observed the environment up to the end of step $t-1$. Its observation $o^{(i)}_t$ is:

$$o^{(i)}_t = \left(p^{(i)}_t,\, p^{(i)}_{\text{init}},\, p^{(j)}_{\text{init}},\, \Delta^{(j)}_{t-1},\, T_t\right),$$

  where $j = 3 - i$. Concretely, Player i sees:

1. **Its own current position** $p_t^{(i)}$.
2. **Its own (true) initial position** $p_{\text{init}}^{(i)}$.
3. **The opponent's initial position** $p_{\text{init}}^{(j)}$.
4. **The opponent's most-recent action** $\Delta_{t-1}^{(j)}$.
5. **The full set of remaining targets** T_t along with their weights $V(\cdot)$. Critically, Player i does *not* observe $p_t^{(j)}$ directly. It must *estimate* $p_t^{(j)}$ using the known initial positions and the history of observed opponent-actions. Formally,

$$\mathcal{O}^{(i)} = W \times W \times W \times \{-1, 0, 1\}^3 \times 2^{T_{\text{init}} \times \mathbb{Z}_{>0}},$$

and

$$O^{(i)}\big(s_t, \Delta_{t-1}^{(1)}, \Delta_{t-1}^{(2)}\big) = \Big(p_t^{(i)}, p_{\text{init}}^{(i)}, p_{\text{init}}^{(j)}, \Delta_{t-1}^{(j)}, T_t\Big).$$

—

6. Belief and Estimation of $p_t^{(j)}$

Since Player i does not directly observe $p_t^{(j)}$, it maintains a *belief* (a distribution) over the possible current positions of j. In principle, at each time t, Player i knows:

1. $p_{\text{init}}^{(j)}$ at $t = 0$.
2. The entire sequence of observed opponent-actions $\{\Delta_0^{(j)}, \Delta_1^{(j)}, \ldots, \Delta_{t-1}^{(j)}\}$ up to the previous step.
3. The deterministic transition rules of the environment.

Hence, Player i can compute exactly

$$\widehat{p}_t^{(j)} = \text{simulate}\big(p_{\text{init}}^{(j)};\ \Delta_0^{(j)}, \Delta_1^{(j)}, \ldots, \Delta_{t-1}^{(j)}\big),$$

where "simulate" means "apply each observed $\Delta^{(j)}$ in turn, clamping/collision-checking against the *estimated* position of Player i in each sub-step." But since Player i also must track its own estimated position (which it knows exactly), there is no stochasticity: Player i can keep a running update for "what Player j must be doing," given that i knows every collision event that j would have experienced. In other words:

- At time $t = 0$, Player i sets $\widehat{p}_0^{(j)} = p_{\text{init}}^{(j)}$.
- For each $t = 0, 1, \ldots$, when $\Delta_t^{(j)}$ becomes known (one step later), Player i does exactly the same "collision + clamp + bounce" computation that the environment would do for Player j at sub-step $t.1$ or $t.2$, using rather:
  1. $p_{\text{est}}^{(j)}$ (previous).
  2. $\Delta_t^{(j)}$.
  3. The *true* position of Player i at the corresponding sub-step (which i knows, since it controls itself). Hence at each step, there is **no actual uncertainty** in $p_t^{(j)}$; it is deterministically reconstructible from the

known initial positions and the observed opponent-actions, together with known turn-order. The only "challenge" is that Player i only learns $\Delta_t^{(j)}$ one sub-step later—still, that is enough to update $\widehat{p}_{t+1}^{(j)}$ exactly. —

7. Complete Game Definition

We now summarize the environment as a two-player, turn-based **partially observable** Markov game with:

1. **State space**

$$\mathcal{S} = \left\{ \,(\,p^{(1)},\, p^{(2)},\, T,\, S^{(1)},\, S^{(2)}\,) \mid p^{(i)} \in W,\; T \subseteq T_{\text{init}},\; S^{(i)} \in \mathbb{Z} \right\}.$$

2. **Action spaces**

$$\mathcal{A}^{(i)} = \{-1,\, 0,\, 1\}^3, \qquad i = 1, 2.$$

3. **Transition function** Deterministic, defined by the two sub-steps (Player 1's move, then Player 2's move) as in Section 3.
4. **Reward functions**

$$R^{(1)}\big(s_t, \Delta_t^{(1)}, \Delta_t^{(2)}, s_{t+1}\big) = r_{t+1}^{(1)} \in \{-1, 0, +V(\tau)\},$$

$$R^{(2)}\big(s_t, \Delta_t^{(1)}, \Delta_t^{(2)}, s_{t+1}\big) = r_{t+1}^{(2)} \in \{-1, 0, +V(\tau)\},$$

as defined in Section 4.
5. **Observation functions**

$$O^{(1)}\big(s_t, \Delta_{t-1}^{(1)}, \Delta_{t-1}^{(2)}\big) = \big(p_t^{(1)},\; p_{\text{init}}^{(1)},\; p_{\text{init}}^{(2)},\; \Delta_{t-1}^{(2)},\; T_t\big),$$

$$O^{(2)}\big(s_t, \Delta_{t-1}^{(1)}, \Delta_{t-1}^{(2)}\big) = \big(p_t^{(2)},\; p_{\text{init}}^{(2)},\; p_{\text{init}}^{(1)},\; \Delta_{t-1}^{(1)},\; T_t\big).$$

Each agent i sees its own current cell, both initial positions, the opponent's last action, and the remaining targets—**but not** $p_t^{(j)}$.
6. **Termination** The episode ends at the first $t+1$ such that either
   - $T_{t+1} = \varnothing$ (all targets are gone), or
   - $t + 1 = T_{\max}$ (if a fixed horizon is imposed).
7. **Discount factor** Typically $\gamma = 1$ for an undiscounted finite horizon, or $\gamma < 1$ otherwise. —

8. Belief Update (Estimating the Opponent's Position)

Although each agent does not directly see the opponent's current cell $p_t^{(j)}$, it knows:

- The true value of $p_{\text{init}}^{(j)}$.
- The entire sequence of observed opponent-actions $\Delta_0^{(j)},\, \Delta_1^{(j)},\, \ldots,\, \Delta_{t-1}^{(j)}$.

- Its own true state and actions, so it knows exactly which collisions or clamps would have affected j. Therefore, each agent can maintain a **deterministic estimate**

$$\widehat{p}_t^{(j)} = \text{UpdatePosition}\big(p_{\text{init}}^{(j)};\ \Delta_0^{(j)}, \Delta_1^{(j)}, \ldots, \Delta_{t-1}^{(j)}\big),$$

where "UpdatePosition" means: apply each $\Delta_k^{(j)}$ in turn,

1. Clamp to $[0..X-1] \times [0..Y-1] \times [0..Z-1]$.
2. If the clamped cell would collide with the *estimated* position of i at that same sub-step, bounce backwards by $-\Delta_k^{(j)}$. Since agent i always knows its own exact position (it controls it), this reconstruction is exact. Hence the environment is deterministic from the vantage of each agent's belief: at time t, agent i knows exactly $\widehat{p}_t^{(j)} = p_t^{(j)}$. —

9. Summary of Key Points

1. **Grid**

$$W = \{0, \ldots, X-1\} \times \{0, \ldots, Y-1\} \times \{0, \ldots, Z-1\}.$$

2. **Initial Positions** Each player i knows both $p_{\text{init}}^{(1)}$ and $p_{\text{init}}^{(2)}$, and those remain fixed.
3. **Turn Order**
   - Sub-step $t.1$: Player 1 chooses $\Delta_t^{(1)}$; environment updates $p^{(1)}, S^{(1)}, T$.
   - Sub-step $t.2$: Player 2 chooses $\Delta_t^{(2)}$; environment updates $p^{(2)}, S^{(2)}, T$.
4. **Collision** If the mover's clamped "forward" position equals the other player's current cell, the mover is bounced backwards by $-\Delta$ (clamped) and receives -1 point. No collection occurs on a backward-bounce.
5. **Targets** Each $\tau_k \in T$ has value $V(\tau_k)$. If a player lands (without collision) on τ_k, that player gains $+V(\tau_k)$ and τ_k is removed from T.
6. **Rewards** At step t, Player 1's reward $r_{t+1}^{(1)} = S_{t+\frac{1}{2}}^{(1)} - S_t^{(1)} \in \{-1, 0, +V(\tau)\}$. Player 2's reward $r_{t+1}^{(2)} = S_{t+1}^{(2)} - S_{t+\frac{1}{2}}^{(2)} \in \{-1, 0, +V(\tau)\}$.
7. **Observation for Player i**

$$o_t^{(i)} = \big(p_t^{(i)},\ p_{\text{init}}^{(i)},\ p_{\text{init}}^{(j)},\ \Delta_{t-1}^{(j)},\ T_t\big).$$

   - Knows its own current position $p_t^{(i)}$.
   - Knows both initial positions $p_{\text{init}}^{(i)}, p_{\text{init}}^{(j)}$.
   - Knows the opponent's last action $\Delta_{t-1}^{(j)}$.
   - Sees all remaining targets T_t with their values.
8. **Belief / Estimation** From these observations, Player i can reconstruct exactly the opponent's current cell $p_t^{(j)}$ by starting from $p_{\text{init}}^{(j)}$ and sequentially applying each observed $\Delta_k^{(j)}$ (with the same "collision-bounce" logic, using i's own true position).

9. **Termination** Episode ends at the first $t+1$ such that $T_{t+1} = \varnothing$ (all targets collected) or $t+1 = T_{\max}$ (if a finite horizon is imposed).
10. **Discount Factor** One typically takes $\gamma = 1$ for an undiscounted episodic setting, or any $\gamma < 1$ otherwise. —

In this formulation, **each agent fully knows**:

- Its own and the opponent's **initial** positions $(p_{\text{init}}^{(i)}, p_{\text{init}}^{(j)})$.
- Its own **current** position $p_t^{(i)}$.
- The opponent's **last** action $\Delta_{t-1}^{(j)}$.
- The set of all **remaining targets** (and their weights).

What an agent **does not see** directly is the opponent's current position $p_t^{(j)}$. However, because it knows:

1. $p_{\text{init}}^{(j)}$;
2. all the opponent's past actions $\Delta_0^{(j)}, \ldots, \Delta_{t-1}^{(j)}$ (revealed one-at-a-time); and
3. its own true positions (so it knows exactly when/where opponent collisions would have occurred),

the agent can reconstruct $p_t^{(j)}$ exactly in a deterministic fashion. In that sense, this is only *partially* observable if you insist that "current opponent position" isn't directly given as part of $o_t^{(i)}$—yet it remains *inferable* from the available information. This completes the formal, math-style definition of the problem under your specified informational assumptions.