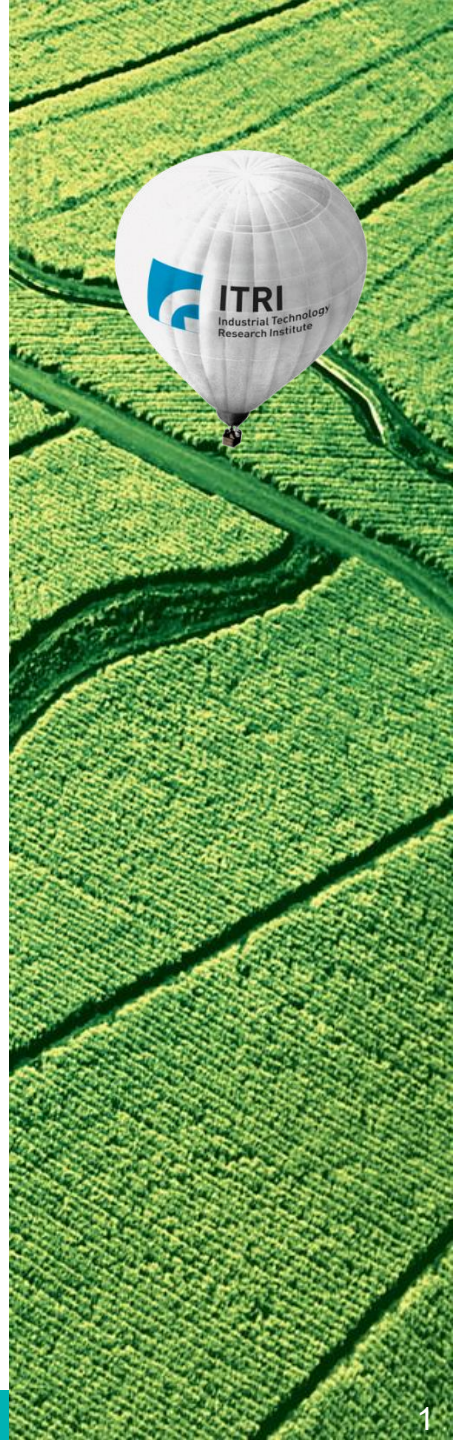
工業技術研究院

Industrial Technology
Research Institute



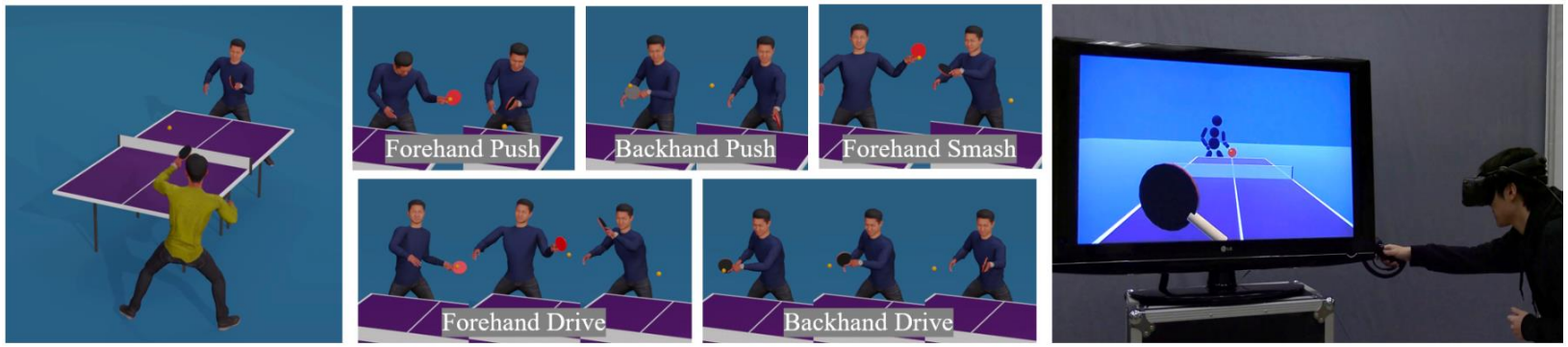
科技藝術書報討論 2025/10/15

Strategy and Skill Learning for Physics-based Table Tennis Animation

SIGGRAPH Conference Papers ’24

陳昱璋 113003856

ABSTRACT



- 近年基於物理的角色動畫結合深度學習後，能產生靈活自然的動作，如後空翻、拳擊與網球等。
  然而，要像人類一樣在動態環境中靈活運用多樣技能仍具挑戰。

- 本研究提出一種**策略與技能學習法**，用於**物理驅動的桌球動畫**。
  此方法解決了「**模式崩塌（mode collapse）**」問題，使角色能更充分發揮多樣運動技能。

- 我們設計了**分層式控制架構**與**策略學習框架**，
  可提升動作自然度與決策效率。

- 透過與最新方法比較及在**AI對AI**與**人對AI**的VR實驗中驗證，
  結果顯示本系統能在競爭與合作情境下，展現更自然且多樣的動作表現。

Jiashun Wang (王家順)

- **學歷與職位：** 博士研究生 (PhD Student)
- **學校： 卡內基梅隆大學 (Carnegie Mellon University, CMU)**
- **學系/學院：** 電腦科學學院 (School of Computer Science) 旗下之 **機器人研究所 (Robotics Institute)**
- **導師：** Jessica Hodgins 教授
- **研究領域：** 他的研究工作主要在**電腦圖形學 (Computer Graphics)**、**電腦視覺 (Computer Vision)** 和**機器人學 (Robotics)** 的交叉領域。他專注於**基於物理的角色動畫 (physics-based character animation)**，同時也關注 3D 人類-機器人-物體互動及其在動畫、元宇宙和機器人技術中的應用。

他先前在加州大學聖地牙哥分校 (UC San Diego) 獲得了碩士學位，並在復旦大學獲得了學士學位。

Publications

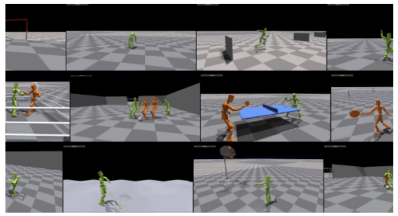
* indicates equal contributions.

ASAP: Aligning Simulation and Real-World Physics for Learning Agile Humanoid Whole-Body Skills
Tairan He*, Jiawei Gao*, Wenli Xiao*, Yuanhang Zhang*, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi "Jim" Fan, Yuke Zhu, Changliu Liu, Guanya Shi
RSS, 2025

paper / project page

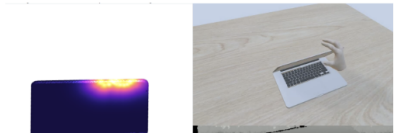
Strategy and Skill Learning for Physics-based Table Tennis Animation
Jiashun Wang, Jessica Hodgins, Jungdam Won
SIGGRAPH, 2024

paper / project page



SMPLOlympics: Sports Environments for Physically Simulated Humanoids
Zhengyi Luo*, **Jiashun Wang***, Kangni Liu*, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, Jessica Hodgins, Kris Kitani
Preprint, 2024

paper / project page



ContactArt: Learning 3D Interaction Priors for Category-level Articulated Object and Hand Poses Estimation
Zehao Zhu*, **Jiashun Wang***, Yuzhe Qin, Deqing Sun, Varun Jampani, Xiaolong Wang

Jessica K. Hodgins

她是電腦圖學與機器人領域中極具代表性的人物之一，
也是這篇論文（*Strategy and Skill Learning for Physics-based Table Tennis Animation*）的第二作者。

- **現任**：卡內基梅隆大學（Carnegie Mellon University, CMU）
電腦科學系（Computer Science Department）與機械工程系（Robotics Institute）教授。
- **過去職務**：
  - 曾任 **Facebook / Meta Reality Labs** 研究副總裁（VP of Research）。
  - 2010–2016 年擔任 **Disney Research Pittsburgh** 研究實驗室負責人。
  - 曾於 **麻省理工學院（MIT）** 任教。

Jungdam Won

•**職位：** 助理教授 (Assistant Professor)
•**學校：** **首爾大學 (Seoul National University)**，位於韓國
•**學系：** 電腦科學與工程學系 (Department of Computer Science and Engineering)
•**研究領域：** 他的研究主要集中在計算機圖形學、機器學習、生物力學、機器人學以及動作分析與合成等領域（他領導了 Intelligent Motion Lab）。
在加入首爾大學任教之前，他曾在 Meta（Facebook）AI 擔任研究科學家。

1. INTRODUCTION（緒論）

研究背景

•深度學習導入**基於物理的角色動畫（Physics-based Character Animation）**，
讓角色能生成**靈活、自然、擬真的動作**。（如後空翻、拳擊、桌球等複雜動作）
•為使角色具備多樣應用性（Versatility），
必須讓技能能在不同環境下重複使用（Reusable Skills）。

現有方法

•多數採兩階段訓練：
① 模仿參考動作 → 學得技能嵌入（Skill Embeddings）
② 任務訓練階段 → 將技能應用於特定任務
•在各種環境下能生成自然動作，但仍有核心問題。

主要挑戰：模式崩塌（Mode Collapse）

•當技能差異細微時，NPC 智能體在訓練後傾向重複使用少數技能。
•導致：

- 忽略學到的多樣性
- 探索受限（Reinforcement Learning exploration reduced）
- 任務表現不佳（Sub-optimal performance）

1. INTRODUCTION（緒論）

研究動機
•以往研究多依賴人工決定技能，智能體**缺乏自動策略選擇能力**。
•本研究提出同時提升**技能控制（Skill Control）與策略決策（Strategy Learning）**的學習架構。

方法概要
分層式技能控制器（Hierarchical Skill Controller）
- 可快速切換桌球技能（forehand, backhand, smash, push...）
- 有效避免 Mode Collapse

策略學習框架（Strategy Learning Framework）
- 讓智能體能根據情境（競爭/合作）自動選擇技能
- 強化人機互動與 AI 決策智能

實驗環境與成果
•**AI 對 AI（Agent-Agent）模擬對戰** → 展現更高技能多樣性與策略性
•**人機互動（Human-Agent in VR）** → 可應對競爭與合作情境

•實驗平台與開源資料：
🔗 https://jiashunwang.github.io/PhysicsPingPong/

2. RELATED WORK
2.1 Physics-based Character Animation — 深度強化學習（DRL）如何改善角色物理控制。

研究背景
•**將物理定律引入角色動畫** → 能產生更真實的行為與動作。【Hodgins 1995; Laszlo 1996】

早期方法
•**軌跡最佳化（Trajectory Optimization）**
→ 找出角色最順暢、最省力的動作路徑。【de Lasa 2010; Mordatch 2012; Yin 2008】
•**取樣法（Sampling-based Methods）**
→ 嘗試許多動作樣本，挑選出最自然的。【Liu 2010, 2016】
問題：這些方法需人工設計「目標函數」，複雜又不靈活。

深度強化學習Deep Reinforcement Learning, DRL的出現
•DRL 大幅提升控制能力與動作自然度。【Liu & Hodgins 2017; Peng 2017】
•優點：
不需人工定義複雜規則
可從大量資料中自行學習控制邏輯
產生流暢自然的角色行為
因此，DRL 成為物理動畫研究的主流技術。

資料驅動式（Data-driven）動畫興起
•**Peng et al. (2018)** 提出 DRL-based 模型後，資料驅動式學習快速普及。
•延伸方向：
- 處理更大資料集【Bergamin 2019; Won 2020】
- 重新組合狀態轉換（動作片段）【Peng 2021】

2. RELATED WORK
2.1 Physics-based Character Animation — 深度強化學習（DRL）如何改善角色物理控制。

可重複使用的運動技能（Reusable Motor Skills）
近期研究聚焦於學習可重複使用的運動技能。
其主要思想是先學習**參考動作的潛在空間（latent space）**，

各類潛在模型包括：
•編碼器 - 解碼器與自回歸模型（Encoder-Decoder with Autoregression）【Merel et al. 2019; Won et al. 2021】
•球面嵌入模型（Spherical Embedding）【Dou et al. 2023; Peng et al. 2022; Tessler et al. 2023】
•條件變分自編碼器（Conditional VAE）【Won et al. 2022; Yao et al. 2022】
•向量量化 VAE（Vector-Quantized VAE）【Zhu et al. 2023】
•局部模型（Part-wise Models）【Bae et al. 2023; Xu et al. 2023】

多智能體角色模擬（Multi-agent Simulation）
本研究的系統設計用於**桌球遊戲（Table Tennis）**，涉及兩名玩家（即兩個智能體，agents）。
過去多智能體的研究主要採用**運動學方法（Kinematic Approaches）**【Kwon et al. 2008; Liu et al. 2006; Shum et al. 2008, 2012; Wampler et al. 2010】。
近期亦出現基於物理模擬的拳擊與網球範例【Won et al. 2021; Zhu et al. 2023; Zhang et al. 2023】。
Zhang 等人（2023）從轉播影片中學習網球動作，
先以運動學方式生成動作，再以物理追蹤（Physics-based Tracking）進行校正，
並利用殘差力與手臂控制以完成擊球。
然而，該方法中的技能與目標選擇並非透過學習，而是由人工或隨機設定。

本研究方法
相較之下，本研究的方法不僅能學習靈活且精確的擊球控制（Motor Control），
亦能根據對手與球的運動，自動學習技能與擊球目標的選擇策略（Skill and Target Selection Strategies）。

2. RELATED WORK
2.2 Transition of Skills — 層級式 RL、Option-based、行為樹等技能轉換方法。

研究背景

•**Option-based Methods**【Bagaria & Konidaris 2020; Jain et al. 2021; Klissarov et al. 2017; Konidaris & Barto 2009; Sutton et al. 1999】

- 將每個技能（Skill）定義為一個「選項（Option）」
- 各選項依序構成動作鏈（Sequence），前一選項的執行結果啟動下一個選項

轉換策略的改進

•**Lee et al. (2019)**

- 提出「**轉換策略（Transition Policies）**」以連接基本技能（Primitive Skills）
- 引入「**鄰近度預測器（Proximity Predictors）**」
  - 根據與下一技能初始狀態的接近程度給予獎勵（Rewards）

•**Lee et al. (2021)**

- 使用「**終端狀態正則化（Terminal State Regularization）**」
  - 確保前後技能狀態連續，用於長期任務（Long-horizon Tasks）

行為樹（Behavior Trees）方法

•廣泛用於規劃狀態轉換【Cheng et al. 2023; French et al. 2019; Marzinotto et al. 2014】

•核心概念：

- 使前一階段的「終端狀態（Terminal State）」接近下一階段的「初始狀態（Initial State）」
- 以確保技能切換過程平滑穩定

桌球任務中的挑戰

•桌球屬於**高速度（High-speed）與高反應性（Rapid-response）**任務

•玩家並非總是從明確定義的初始狀態擊球

•傳統技能轉換方法難以應對此類即時動態任務

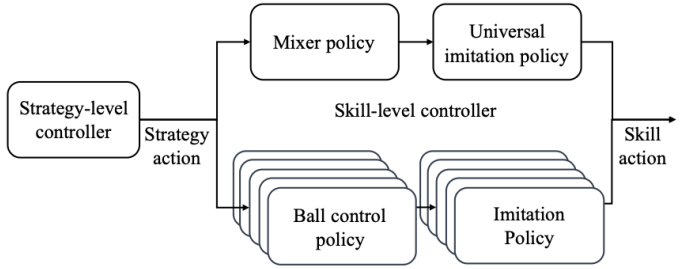


Figure 2: An overview of our method. Strategy action includes the skill command and ball's target landing location. Skill action includes the target joint angles for PD controllers, blended from the outputs of imitation policies.

本研究的動作決策方法，出現在這裡太早了

2. RELATED WORK
2.3 Human-Agent Interaction — VR 環境中人機互動的研究背景與技術挑戰。

研究背景

•過去研究多聚焦於 **VR 運動訓練系統**【Liu et al. 2020; Pastel et al. 2023】
→ 但**缺乏完整物理模擬的對手（Physically Simulated Opponent）**
•商業化應用：
- 常見於 **拳擊、高爾夫、羽球** 等 VR 遊戲
- 代表例：**Eleven Table Tennis (2016)**
  → 玩家可與 AI 對打
  → 但智能體僅模擬「漂浮的頭部與球拍」，**非全身動態模擬**

技術進展

•隨著 **GPU 加速模擬** 與 **控制演算法** 的進步
→ 本研究能建立具 **全身物理動力學（Full-body Dynamics）** 的智能體
→ 使其可在 VR 中與人類**即時對打（Real-time Play）**

延伸研究方向

•「**人類在迴路中（Human-in-the-loop）**」與 **延展實境（XR）** 的方法用於增強智能體的學習與互動能力
【Brenneis et al. 2021; Li et al. 2022; Seo et al. 2023; Wang et al. 2023】

本研究的突破（Contribution）

我們將 **人類與智能體** 整合進同一個
統一的物理模擬環境（Unified Physical Environment），
使雙方能進行 **雙向物理互動（Bidirectional Physical Interaction）**，
並能同時展現出 **合作（Cooperation）與競爭（Competition）** 行為。

3. METHOD OVERVIEW

我們提出了一種**階層式架構（Hierarchical Approach）**，
其包含兩個主要的控制層級：
策略層控制器（Strategy-level Controller）
技能層控制器（Skill-level Controller）

策略層控制器（Strategy-level Controller）
此控制器將以下狀態作為輸入：
•智能體自身的狀態（state of the agent）
•對手的狀態（state of the opponent）
•球的狀態（state of the ball）
然後輸出一個「策略動作（strategy action）」，
該動作包含：
•要使用的技能（the skill to use）
•球的預期落點（the target landing location for the ball）

技能層控制器（Skill-level Controller）
技能層控制器的輸入包括：
•智能體與球的狀態（the states of the agent and ball）
•從策略層控制器輸出的策略動作（the strategy action）

此控制器接著產生「技能動作（skill action）」，其中包含：
•供 **PD 控制器（Proportional-Derivative Controllers）**使用的關節目標角度（target joint angles）

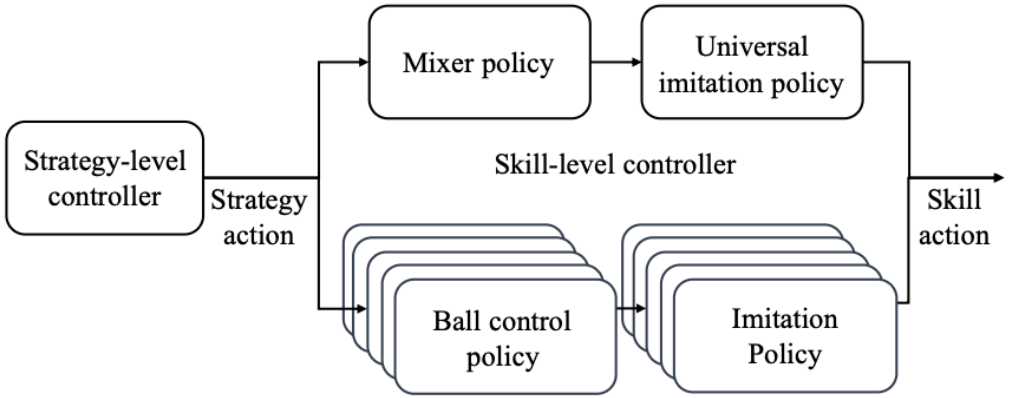


Figure 2: An overview of our method. Strategy action includes the skill command and ball's target landing location. Skill action includes the target joint angles for PD controllers, blended from the outputs of imitation policies.

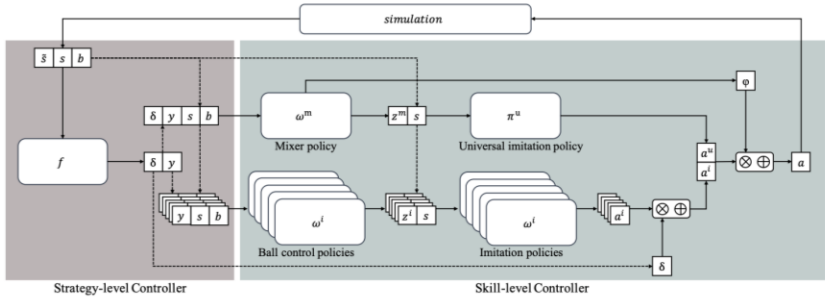


Figure 3: The architecture of our method. We train the skill-level controller through the stages of imitation policies, ball control policies, and finally, the mixer policy. We train the strategy-level controller after the skill-level controller is ready and its weight is frozen. ⊗⊕ stands for the weighted sum in Equation 8.

4. SKILL-LEVEL CONTROLLER（技能層控制器）

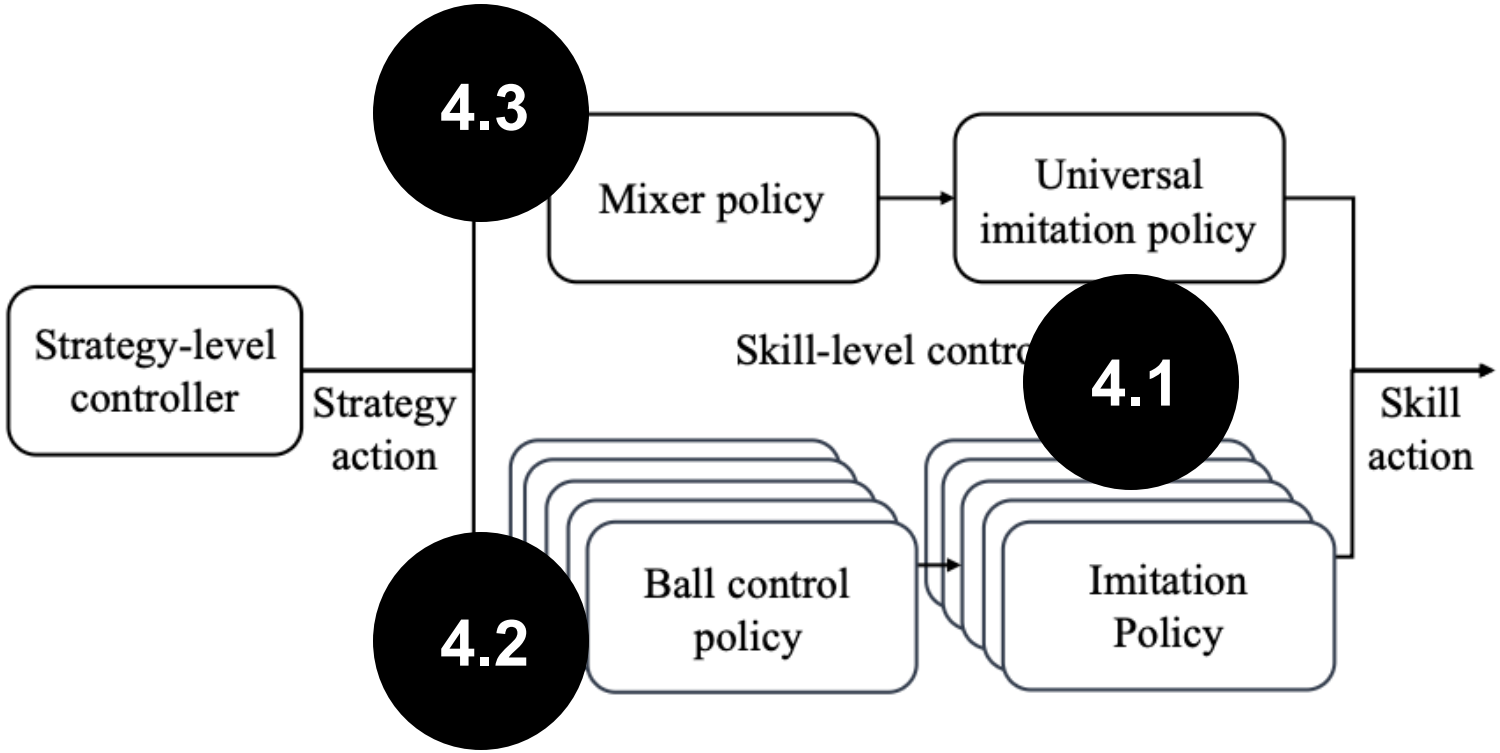


Figure 2: An overview of our method. Strategy action includes the skill command and ball's target landing location. Skill action includes the target joint angles for PD controllers, blended from the outputs of imitation policies.

4. SKILL-LEVEL CONTROLLER（技能層控制器）

三階段訓練架構

4.1 模仿階段（Imitation Stage）

使用動作擷取資料（Motion Capture Data）
訓練出多個模仿策略（Imitation Policies），
讓 AI 能再現真實擊球動作。

4.2 控球階段（Ball Control Stage）

為每一種技能建立對應的「球控策略（Ball Control Policy）」
→ 讓智能體能根據模仿策略回擊不同來球。

4.3 混合階段（Mixer Stage）

訓練「混合策略（Mixer Policy）」
→ 讓智能體能**自然地連接不同技能**，
實現連續、流暢、合理的動作轉換。

當整個技能層控制器訓練完成後：
•智能體能**流暢連續地執行多種技能**
•並能**精準地將球擊向多個不同目標位置**

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•4.1 Imitation Policy — 使用動作捕捉（Motion Capture）資料訓練模仿。

研究目的

讓 AI 能「學會打桌球」，在物理模擬環境中學會各種擊球技能。

資料與模型設計

•將動作資料分成五種技能（正手、反手、推擋、扣殺等）。
•為每個技能訓練一個模仿策略 π_1–π_5。
•用全部資料再訓練一個「通用模仿策略」 π_u。

模型輸入與輸出：

•**s**：AI 狀態（姿勢、位置等）
•**z**：潛在變數（動作風格）
•**a**：AI 執行的動作（關節角度）

四個公式對應的訓練重點

$$\min_{D^i} -\mathbb{E}_{d_{M^i}(s,s')}\log(D^i(s,s')) - \mathbb{E}_{d_{\pi^i}(s,s')}\log(1-D^i(s,s')) + \lambda_{gp}\mathbb{E}_{d_{M^i}(s,s')}\left\|\nabla_\phi D^i(\phi)\big|_{\phi=(s,s')}\right\|^2, \quad (1)$$

(1) 對抗訓練 — 讓 AI 動作像人

AI（模仿模型）與鑑別器 D_n 對抗學習，
逼近真實動作分佈，學會「以假亂真」。
→ 提升動作真實性

$$\max_{q^i}\mathbb{E}_{p(z^i)}\mathbb{E}_{d^{\pi^i}(s,s'|z^i)}[\log q^i(z^i|s,s')], \quad q^i(z^i|s,s') = \frac{1}{Z}exp(\mu_{q^i}(s,s')^T z^i) \quad (2)$$

(2) 特徵學習 — 建立 z 與動作對應

訓練編碼器 q_n，
讓潛在變數 **z** 能對應到具體的動作風格（如正手推、反手拉）。
→ 提升 AI 的動作辨識與控制能力

$$r_t = -\log(1-D^i(s_t,s_{t+1})) + \beta\log q^i(z_t^i|s_t,s_{t+1}). \quad (3)$$

(3) 獎勵設計 — 鼓勵真實又穩定的動作

結合兩個訊號：
•動作越逼真，獎勵越高
•潛在變數與動作轉換越一致，獎勵越高
→ 提升穩定性與可解釋性

$$\max_{\pi^i}\mathbb{E}_{p(Z)}\mathbb{E}_{p(\tau|\pi^i,Z)}[\sum_{t=0}^{T}\gamma^t(r_t)] - \lambda_{D^i}\mathbb{E}_{d^{\pi^i}(s)}\mathbb{E}_{z_1^i,z_2^i\sim p(z^i)}\left[\left(\frac{D_{KL}(\pi^i(\cdot|s,z_1^i),\pi^i(\cdot|s,z_2^i))}{0.5(1-z_1^iz_2^i)}-1\right)^2\right], \quad (4)$$

(4) 整體優化 — 兼顧真實性與多樣性

在整體訓練中同時最大化獎勵、最小化模式崩塌。
確保 AI 不會只用一招吃遍天。
→ 提升技能多樣性與靈活度

一句話總結

這四個公式構成一個完整閉環：

模仿 → 判斷 → 給獎勵 → 維持多樣性。

讓 AI 不只會「模仿人」，還能「像人一樣打球」。

階段	對應公式	功能	比喻說明
對抗階段	公式 (1)	「辨真偽」— 讓動作看起來像真人	鑑別器在打分，看 AI 動作像不像人
特徵學習階段	公式 (2)	「學習特徵」— 建立 z 與動作變化的對應	學會把不同的動作風格壓縮成 z 向量
獎勵學習階段	公式 (3)	給予訓練訊號，鼓勵真實與多樣	結合對抗分數與潛在一致性分數
最終整合階段	公式 (4)	綜合優化 — 同時追求逼真與多樣	讓 AI 在不同動作下都能表現自然

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•**4.1 Imitation Policy** — 使用動作捕捉資料訓練模仿政策。

$$\min_{D^i} -\mathbb{E}_{d_{M^i}(s,s')} \log(D^i(s,s')) - \mathbb{E}_{d_{\pi^i}(s,s')} \log(1 - D^i(s,s'))$$
$$+ \lambda_{gp} \mathbb{E}_{d_{M^i}(s,s')} \left\| \nabla_\phi D^i(\phi)\big|_{\phi=(s,s')} \right\|^2, \qquad (1)$$

公式 (1)：對抗損失（Adversarial Loss）

目的：

讓「AI 生成的動作」騙過「鑑別者（Discriminator）」

→ 讓模擬動作與真實動作越像越好。

概念比喻：

就像 GAN（生成對抗網路），AI 嘗試「模仿高手」，而鑑別器負責分辨真假，兩者互相競爭直到分不出差別。

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•**4.1 Imitation Policy** — 使用動作捕捉資料訓練模仿政策。

$$\max_{q^i} \mathbb{E}_{p(z^i)} \mathbb{E}_{d^{\pi^i}(s,s'|z^i)} [\log q^i(z^i|s,s')], \tag{2}$$

$$q^i(z^i|s,s') = \frac{1}{Z} exp(\mu_{q^i}(s,s')^T z^i)$$

公式 (2)：編碼器訓練（Latent Encoder Training）

目的：

建立「潛在變數」ziz_izi 與「動作轉換 (s→s')」之間的對應關係。

意思：

讓每一個隱藏向量 ziz_izi 都代表一種具體動作特徵，

例如「正手推」、「反手拉」等。

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•**4.1 Imitation Policy** — 使用動作捕捉資料訓練模仿政策。

$$r_t = -\log(1 - D^i(s_t, s_{t+1})) + \beta \log q^i(z_t^i | s_t, s_{t+1}). \qquad (3)$$

公式 (3)：獎勵函數（Reward Function）

目的：

結合「真實度獎勵」與「潛在一致性獎勵」。

兩部分：

•第一項：動作越像真實人類 → 獎勵越高。

•第二項：潛在變數能正確描述動作 → 獎勵越高。

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•**4.1 Imitation Policy** — 使用動作捕捉資料訓練模仿政策。

$$\max_{\pi^i} \mathbb{E}_{p(Z)} \mathbb{E}_{p(\tau|\pi^i,Z)} [\sum_{t=0}^{T} \gamma^t (r_t)] \tag{4}$$

$$- \lambda_{D^i} \mathbb{E}_{d^{\pi^i}(s)} \mathbb{E}_{z_1^i, z_2^i \sim p(z^i)} \left[\left(\frac{D_{KL}(\pi^i(\cdot|s, z_1^i), \pi^i(\cdot|s, z_2^i))}{0.5(1 - z_1^i z_2^i)} - 1 \right)^2 \right],$$

公式 (4)：最終目標函數（Policy Objective）

目的：

在整體訓練中最大化總獎勵，同時保持動作多樣性。

重點：

•前半：強化學習式地累積獎勵。

•後半：加上「多樣性懲罰」— 避免所有潛在變數都產生一樣的動作。

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•4.2 Ball Control Policy — 訓練 AI 控制球拍與落點。

讓 AI 學會：
「把隨機來球打回指定落點，動作又自然。」

模型結構

球控策略：

- **s**：智能體狀態（姿勢、位置、速度）
- **b**：球的狀態（位置、速度）
- **y**：目標落點（Target landing location）
- **z_i**：技能潛在變數（Skill latent code）

$$\omega_i(z_i|s,b,y)$$

類別	名稱	意義
r_p	Paddle Reward	球拍靠近球越近 → 獎勵越高
r_β	Ball Reward	球飛向正確目標 → 獎勵越高
r_s	Style Reward	動作越自然 → 獎勵越高

$$r(t) = w_p r_p(t) + w_b r_b(t) + w_s r_s(t)$$

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•4.2 Ball Control Policy — 訓練 AI 控制球拍與落點。

三項獎勵函數

(1) Paddle Reward 球拍獎勵

球拍越靠近球 → 獎勵越高

（$C_{bp}=0$ 代表尚未接觸）

$$r_p(t) = \begin{cases} \exp(-4||x_p(t) - x_b(t)||^2), & \text{若 } C_{bp}(t) = 0 \\ 0, & \text{否則} \end{cases}$$

(2) Ball Reward 球體獎勵

球被正確打向目標落點 → 獎勵最高

$$r_b(t) = \begin{cases} 1 + \exp(-4||x_c(t) - x_t(t)||^2), & \text{若 } C_{bp}(t) = 1 \text{ 且 } C_{bt}(t) = 0 \\ 0, & \text{否則} \end{cases}$$

(3) Style Reward 動作風格獎勵

由「鑑別器 D_i」評估動作是否自然，越像真人 → 獎勵越高。

$$r_s = -\log(1 - D_i(s_t, s_{t+1}))$$

小結

•三種獎勵協同作用：

→ **準確（Accuracy）+ 流暢（Smoothness）+ 真實（Naturalness）**

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•4.3 Mixer Policy — 混合多技能控制器，實現自然技能轉換並避免 mode collapse。

•AI 已能透過「球控策略（Ball Control Policy）」搭配模仿策略（Imitation Policy）進行擊球。
•**限制**：每次只能使用單一技能（如正手、反手…）。
•若直接在比賽中切換控制器，容易失敗。
→ 因為上一個技能的**結束狀態**與下一個技能的**起始狀態**不連貫。

解決方案：混合策略（Mixer Policy）
建立一個新的控制器：

輸入：
•**s**：智能體狀態（Agent State）
•**b**：球的狀態（Ball State）
•**δ**：技能選擇（One-hot 向量，表示要使用哪種擊球技能）
•**y**：球的目標落點（Target Landing Location）

$$\omega_m(z_m|s,b,\delta,y)$$

輸出：
潛在變數 $\mathbf{z_m}$ → 給通用模仿策略（π_u）
混合權重 **φ (phi)** → 控制各技能融合比例
功能概念
•φ 會決定智能體在技能轉換時「依賴誰」：
- φ → 越大 → 越傾向通用策略（π_u）
- φ → 越小 → 越傾向特定技能策略（π_1–π_5）

•透過這樣的**關節層級融合（Joint-wise Blending）**，
AI 可以在不同技能間「平滑切換」，實現連貫的擊球行為。

工業技術研究院
Industrial Technology Research Institute

4. SKILL-LEVEL CONTROLLER（技能層控制器）

•**4.3 Mixer Policy** — 混合多技能控制器，實現自然技能轉換並避免 mode collapse。

$$a = \varphi \odot \pi_u(\cdot \mid s, z_u) + (1 - \varphi) \odot \sum_{i=1}^{5} \delta_i \pi_i(\cdot \mid s, z_i)$$

符號	意義
a	目標關節角度（Target joint angles）
π_u	通用模仿策略（Universal Imitation Policy）
π_i	各技能對應的模仿策略（Skill-specific Imitation Policies）
δ_i	One-hot 向量，指定哪個技能被啟用
φ	混合權重，控制各策略融合比例

效果

AI 能平滑切換技能（例如從防守轉為進攻）。
動作過渡自然、不出現斷層。
支援多技能連續打擊（multi-skill combo）。

5. STRATEGY-LEVEL CONTROLLER（策略層控制器）

讓 AI 能在打球時「**自己想策略**」
決定：
用哪個技能（Skill）
把球打到哪裡（Target）

輸入與輸出
•**輸入**：AI 狀態 s、對手狀態 $\tilde{s}$、球狀態 b
•**輸出**：技能指令 δ、落點目標 y
訓練方式：行為模仿（Behavior Cloning）

模型核心：CVAE
•**Encoder**：把情境與策略壓縮成潛在向量 **z (或 u)**
•**Decoder**：用 (情境 + z) 生成新的策略行為
•**z** 代表策略的「隱藏基因」，控制打法風格

訓練目標
模仿專家策略
保持潛在空間平滑、多樣

重建損失（Reconstruction Loss）
•功能：讓模型**學會模仿人類或高水準 AI 的決策**。

$$L = \sum_k ||c'_k - c_{expert,k}|| + \beta_{KL} D_{KL}(Q(u|\mu_k, \sigma_k^2)||N(0,I))$$

第二項：KL 散度損失（KL Divergence Loss）
•功能：
- 保持 **潛在空間（latent space）** 平滑、連續。
- 讓模型能夠「隨機抽樣 u（或 z）」生成多樣化策略。
- **u（或 z）** 就是「隨機但合理的策略風格」。

工業技術研究院
Industrial Technology Research Institute

6. INTERACTION ENVIRONMENT（互動環境）

驗證我們提出的 **策略學習方法（Strategy Learning Approach）**，
透過兩種互動環境：
Agent–Agent Interaction（AI 對 AI）
Human–Agent Interaction（人對 AI）

系統設計概述
•使用 **物理模擬環境（Physics Simulation）**：Isaac Gym
•視覺呈現透過 **Unity + VR 裝置**（HMD + 手持控制器）

類型	說明	主要目的
Agent–Agent	兩個虛擬角色互打桌球	驗證策略學習與競爭/合作效果
Human–Agent	人類使用 VR 與 AI 對打	驗證 AI 對人類互動的即時反應與學習能力

6. INTERACTION ENVIRONMENT（互動環境）

Agent–Agent Interaction

雙智能體互動架構
•兩個 AI 角色互相比賽：
一個是「**我們的 Agent**」，另一個是「**對手（Opponent）**」。
•透過不斷對戰與資料收集，**迭代訓練策略控制器（Strategy-Level Controller）**。

類型	定義	行為特徵	測試目的
Random Strategy Opponent	隨機選擇技能與落點	動作不連貫、隨機性高	測試穩定性與容錯力
Video Strategy Opponent	來自 20 分鐘真實轉播影片訓練	動作自然、模擬人類打法	測試策略與競爭能力

學習方式
•**競爭模式（Competition）**：選取「獲勝」回合資料作為學習樣本。
•**合作模式（Cooperation）**：選取「成功回球」的長回合資料作為學習樣本。
•每次訓練後更新策略控制器，形成「自我博弈式（Iterative Self-play）」強化。

6. INTERACTION ENVIRONMENT（互動環境）

VR 互動環境
•使用者戴上 **頭戴顯示器（HMD）**，手持 **VR 控制器** 充當球拍。
•控制器的空間位置（q_user）會傳入模擬環境，

實現效果
•**AI 與人類的即時物理互動**（Bidirectional Physical Interaction）。
•Unity 負責即時渲染，Isaac Gym 負責模擬物理運動。
•資料傳輸量大幅降低，可實現**真實打球體驗**。

結果應用
•模型可從人類互動資料中進一步學習策略。
•使用與 Agent–Agent 相同的（Pipeline）訓練策略控制器。
•成為日後「人–AI 共訓練（Co-Learning）」的基礎環境。

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

- 評估 AI「桌球選手」的表現，從兩個面向：
  動作品質（Motion Quality）：
  動作自然嗎？有照指令打對招嗎？
- **任務表現（Task Performance）**：
  打得準不準？能連續回幾球？

模型	說明
ASE (Peng et al., 2022)	傳統對抗式動作模仿模型
CASE (Dou et al., 2023)	加入球體控制與姿態預測
ET	我們的方法去掉 Mixer Policy 的變體
Ours	含完整分層控制器（Skill + Mixer + Strategy）

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

動作品質（Motion Quality）

指標	說明	意義
Discriminator Score	動作像真人的程度	分數越高越自然
Skill Accuracy	打對技能的比例	準確率越高越聰明
Diversity Score	各技能差異性	分數越高越多樣

Table 1: Comparisons on *Discriminator Score*, *Skill Accuracy*, and *Diversity Score*.

	ASE	CASE	ET	Ours
Discriminator Score	1.62	2.28	4.95	**5.72**
Skill Accuracy	0.38	0.47	0.69	**0.76**
Diversity Score	6.13	6.05	7.32	**8.01**

- 我們的方法生成的動作最自然（+15.6%）
- 擊球技能判斷最準確（76% 正確率）
- 動作多樣性最高（比 ASE 高 30%）

工業技術研究院
Industrial Technology Research Institute

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

任務表現（Task Performance）

持續性（Sustainability）
→ 平均能連續回幾球
準確度（Accuracy）
→ 球實際落點與目標落點距離

Table 2: Task performance evaluation. Our method can achieve the longest average hits and the second best accuracy.

	ASE	CASE	ET	Ours
Avg Hits	9.54 (5.94)	8.79 (5.28	6.55 (3.66)	**10.93 (6.28)**
Avg error	0.28 (0.33)	0.35 (0.39)	**0.25 (0.28)**	0.26 (0.31)

- 我們的模型最耐打（平均 10.9 次）
- 準確度排名第二（僅次於 ET）
- ET 雖準但反應慢 → 維持性差

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

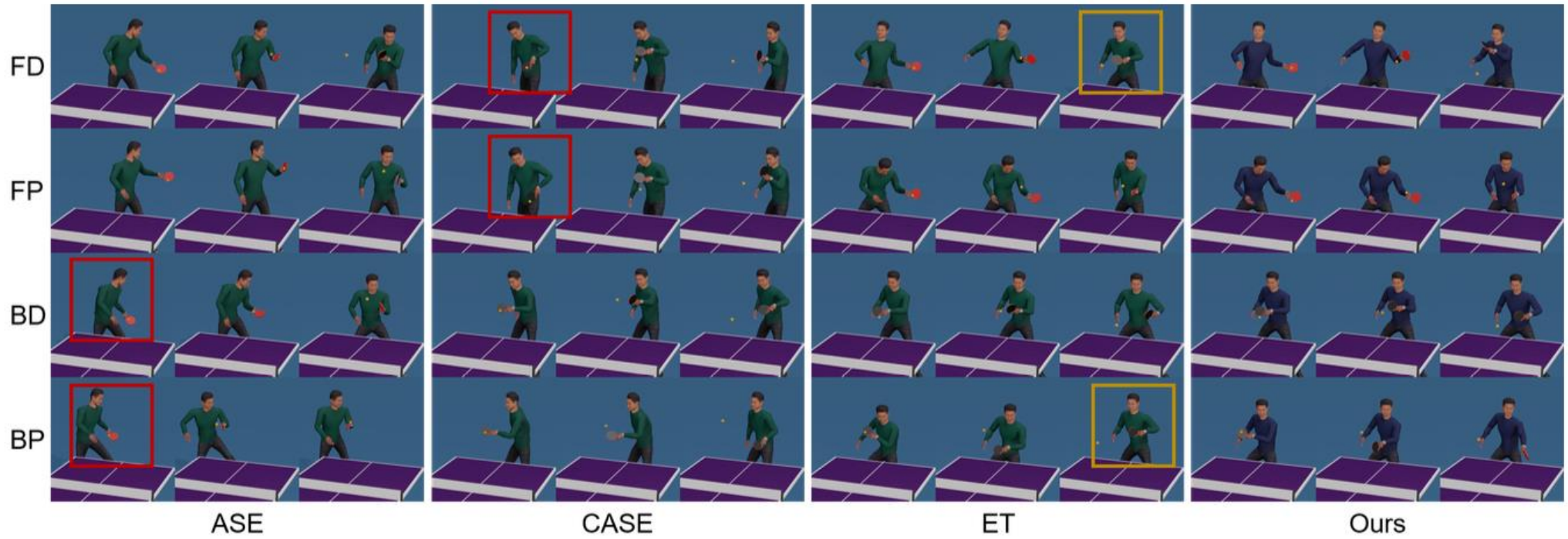


Figure 4: Comparison with other methods with four skill commands. ASE and CASE may use wrong skills as shown in the red box. ET may terminate earlier to return to a preparation pose, as shown in the yellow boxes.

（ASE、CASE、ET、Ours）在相同擊球指令下的動作比較圖
- 紅框：ASE、CASE 打錯技能（例如 應該反手卻用了正手）
- 黃框：ET 提早結束動作（沒完整打完）
- 我們的模型 → 動作連貫且正確

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

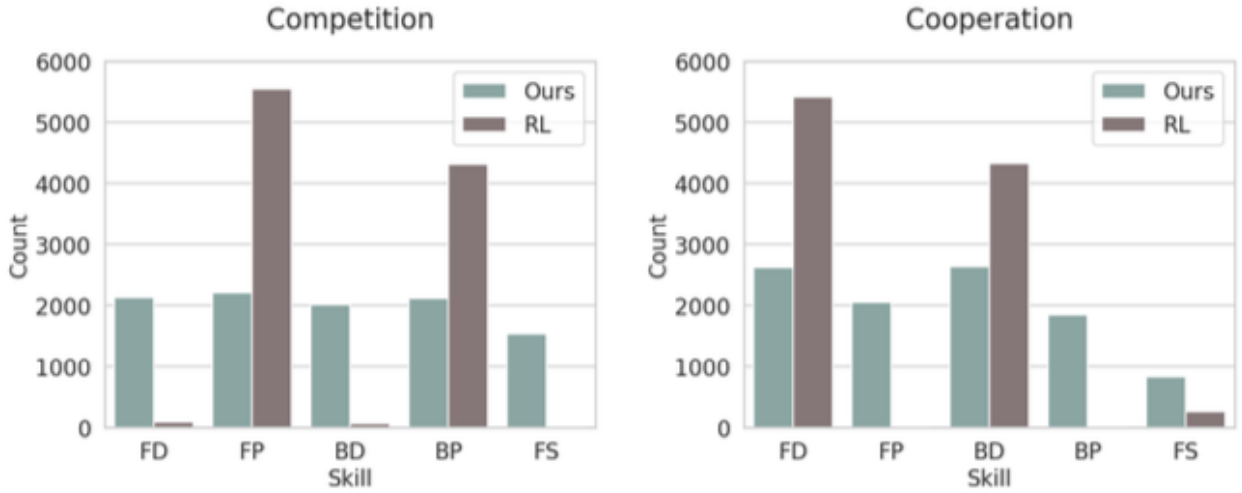


Figure 6: Skill command distribution of our method and RL.

•我們的方法與一般 **RL (Reinforcement Learning)** 模型在比賽過程中所使用的技能指令分布。
•換句話說，它在展示「AI 選擇技能的傾向」。

內容重點：
•Ours（本研究）：技能使用分布平均、四種技能（FD / FP / BD / BP）皆會使用。
•RL baseline：偏好重複少數技能（例如只用 Forehand Drive），顯示出 **Mode Collapse** 現象。
用途：
→ 視覺化地證明我們的方法能在多種技能間自由切換，
而非只重複固定動作。

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

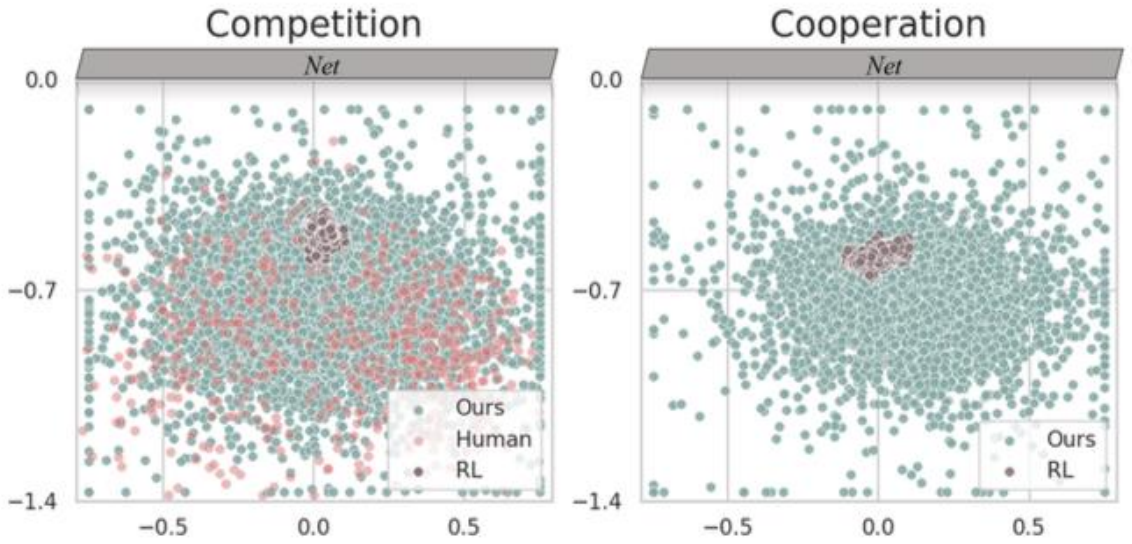


Figure 7: Target landing locations of our method, RL and Human.

- **Human（人類選手）**：落點分布廣、策略多變。
- **RL baseline**：落點集中在固定區域（單一策略）。
- **Ours（本研究）**：落點分布與人類相似，顯示出**策略多樣性與靈活性**。

用途：
→ 用視覺證據支持 Table 2 的任務表現評估，
表明我們的方法能同時達成「準確落點」與「多樣策略」。

7. EXPERIMENTS
7.1 Skill Evaluation — 評估動作自然度與任務表現（Discriminator Score, Skill Accuracy, Diversity)

混合策略（Mixer Policy）觀察
我們觀察不同技能下的混合權重（φ）：
•擊球瞬間 → φ 最低（靠 Ball Control）
•換招階段 → φ 提高（靠 Mixer Policy）

結論
•混合策略能「柔順地」在技能間轉換
•不會硬切換造成動作中斷
•關節（肩、肘、腕）的權重會自然變化 → 打球更像人

我們的 AI 球員：
•動作自然（像人）
•技能正確（會變招）
•多樣穩定（連打不錯過）

Mixer Policy + 分層控制結構
→ 成功避免 Mode Collapse，達成自然連續的多技能運動表現。

7. EXPERIMENTS
7.2 Agent–Agent Interaction Evaluation — 檢驗策略學習成效與勝率。

評估策略學習（Strategy Learning）在兩種情境下的表現：
競爭模式（Competition）：AI 嘗試擊敗對手。
合作模式（Cooperation）：AI 嘗試維持更長回合數。

比較基準
作為基準模型，研究者同時訓練了一個傳統
強化學習策略（RL baseline）

以比較兩者在不同環境下的表現。
對手類型（Opponent Types）
來自第六章的兩種對手設定：

類型	名稱	說明
Random Strategy	隨機選擇技能與落點（無規則）。	
Video Strategy	從真實桌球比賽影片擷取動作資料並訓練成策略控制器。	

工業技術研究院
Industrial Technology Research Institute

7. EXPERIMENTS
7.2 Agent–Agent Interaction Evaluation — 檢驗策略學習成效與勝率。

Table 3: Strategy evaluation. We report the winning rates for the competition setting and average rounds for the cooperation setting.

	Competition		Cooperation	
	RL	Ours	RL	Ours
Random op	0.641	**0.687**	14.9	**16.4**
Video op	0.637	**0.681**	15.6	**18.2**

結果重點

- 我們的策略學習在 **競爭模式勝率更高**。
- 在 **合作模式中能延長回合數**（維持更流暢互動）。
- 對兩種對手類型皆表現穩定，展現出策略適應能力。

7. EXPERIMENTS

7.2 Agent–Agent Interaction Evaluation — 檢驗策略學習成效與勝率。

Table 4: Winning rates between our method and RL. The opponent in parentheses is the opponent during training of the strategy policy.

	Ours (random op)	Ours (video op)
RL (random op)	0.45 vs 0.55	0.47 vs 0.53
RL (video op)	0.42 vs 0.58	0.42 vs 0.58

全面勝出

我們的方法在四場對戰中全部勝出（勝率皆 > 0.5）。

泛化性更好

無論對手是「隨機策略」或「影片策略」，

Ours 都能維持穩定表現。

→ 說明模型沒有「過度學特定對手」（overfitting）。

7. EXPERIMENTS
7.2 Agent–Agent Interaction Evaluation — 檢驗策略學習成效與勝率。

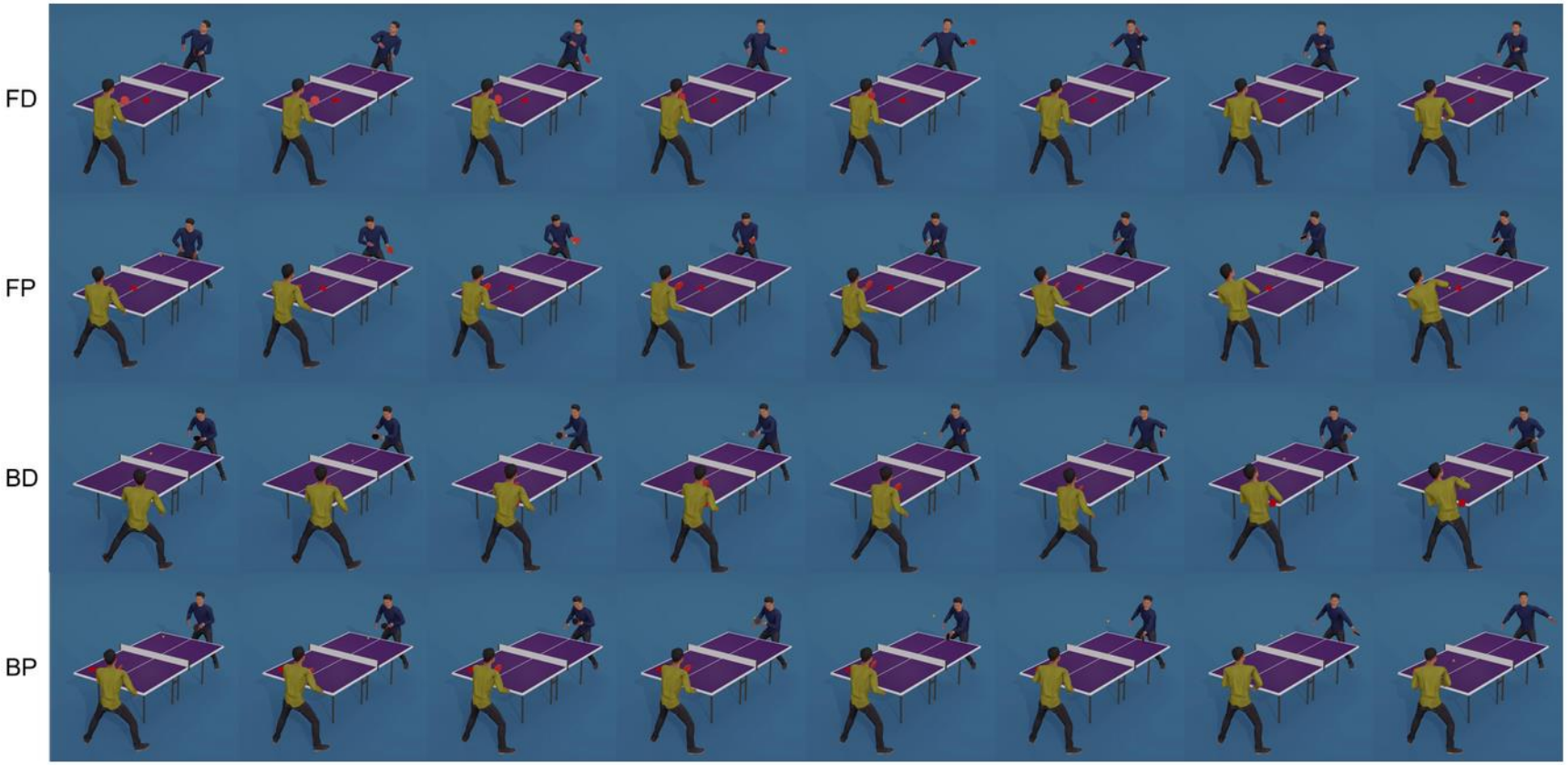


Figure 9: Agent-agent gameplay. Blue agent is applying our strategy-level controller. The red dot is the target. We demonstrate four skills; the forehand smash is less obvious because the opponent does not deliver high and slow shots.

7. EXPERIMENTS
7.3 Human–Agent Interaction Evaluation — VR 對戰中之表現與人類互動反應。

驗證 AI 桌球選手是否能與真實人類對戰，
並根據人類表現自我調整策略。

領域差距（Domain Gap）

人類在 VR 裡的擊球方式、反應節奏
與模擬環境中的虛擬角色不同，
因此需要「微調（Finetuning）」技能控制器。

Finetuning 的目的：
讓 AI 的揮拍風格更貼近真人打球節奏。

訓練流程

1.先用人類在 VR 中與 AI 對打的資料微調技能層。
2.讓 AI 根據不同策略學習：
    1. **競爭策略 (Competition)**：學會進攻、追求勝利。
    2. **合作策略 (Cooperation)**：學會放慢節奏、維持來回。
3.每次對打的資料都會用來更新 AI 的策略決策模型。

7. EXPERIMENTS

7.3 Human–Agent Interaction Evaluation — VR 對戰中之表現與人類互動反應。

- 競爭策略學會「更強的回擊」→ 勝率上升至 **78%**
但由於攻擊性提高，回合數略減。
- 合作策略學會「延長來回」→ 平均回合提升至 **5.34**
但主動讓分，勝率下降至 **58%**。

Table 5: Evaluation of human-agent interaction.

	Initial policies	Competition	Cooperation
Winning rate	0.64	0.78	0.58
Avg hits	4.04	3.75	5.34

- AI 透過「微調 + 策略學習」能針對真人玩家調整行為風格。
- 證明本研究的 **策略層學習方法** 也適用於真實人機互動環境。

7. EXPERIMENTS
7.3 Human–Agent Interaction Evaluation — VR 對戰中之表現與人類互動反應。

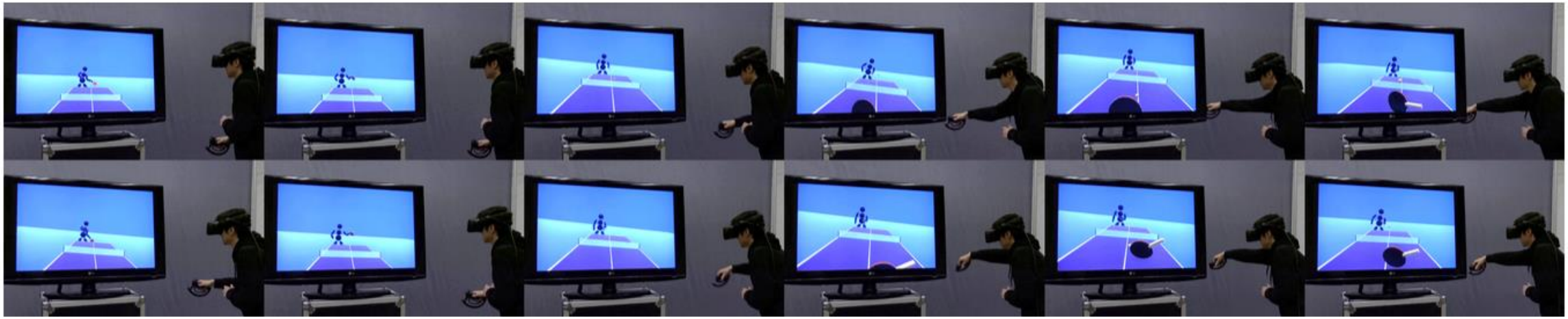


Figure 10: Human-agent interaction screenshots. A human controls a simulated paddle and the agent is simulated and controlled by our method.

8. DISCUSSION AND CONCLUSION（討論與結論）

研究成果
•產生自然、靈活、具競爭力的虛擬桌球智能體。
•採 **階層式控制架構**：
- Skill-Level：負責動作生成與切換。
- Strategy-Level：負責決策與策略。

•有效解決「模式崩塌（Mode Collapse）」問題。
•可於 **AI 對 AI**、**AI 對人類** 兩種環境中運作。

研究限制與未來方向
擴展性（Scalability）
•多技能結構難以擴展到上百種技能。
•未來：結合未標註資料的混合式模型。

資料依賴（Data Dependency）
•動作品質高度影響最終生成結果。
•AI 會模仿玩家習慣，如揮臂過大。

模擬簡化（Physical Simplification）
•未納入「馬格努斯效應（Magnus Effect）」。
•球軌跡略不真實，影響策略學習。

總結
成功建立能學習策略與技能的物理模擬智能體。
驗證於多環境（AI–AI、AI–Human）。
奠定智慧運動與人機互動模擬的研究基礎。

Thank You

本研究的一部分工作，是在 **Jiashun Wang（王嘉順）** 於 **The AI Institute（AI 研究院）** 實習期間完成。

Jungdam Won 的研究部分，
獲得韓國政府（**MSIT**，科學技術資訊通信部）
所資助的 **IITP（資訊通信技術規劃與評估院）** 計畫支持，

計畫名稱為：
「人工智慧研究生院計畫（Artificial Intelligence Graduate School Program）」
（首爾大學，計畫編號：**NO.2021-0-01343-004**），
以及首爾大學 **ICT（電腦技術研究所）** 的部分資助。

我們特別感謝 **Murphy Wonsick** 協助建立 **VR 系統**，
以及 **Melanie Danver** 協助進行 **成果渲染（rendering）**。