UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Arrocha Gascón, Óscar. *Characterizing human state space representations with Deep Reinforcement Learning Models*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Intel·ligència Artificial)

This version is available at https://ddd.uab.cat/record/336300

under the terms of the  license.

# Characterizing human state space representations with Deep Reinforcement Learning Models

Óscar Arrocha Gascón

June 19, 2026

**Abstract**

This project investigates the degree of aligment alignment between human perceptual representations of Atari game states and the internal representations learned by a Deep Q-Network (DQN) across its processing hierarchy, with the hypothesis that human similarity judgments correspond most strongly with the fully connected layer. Human representations are collected through two online behavioral experiments based on a triplet odd-one-out paradigm and compared with DQN layer activations using Representational Similarity Analysis (RSA) across three games: Pong, Ms. Pac-Man, and Space Invaders. The results reveal a heterogeneous picture: partial human–DQN alignment is found in Pong, localized to the fully connected layer; strong layer-independent alignment is observed in Space Invaders; and no reliable alignment is detected in Ms. Pac-Man.

**Keywords:** Deep Reinforcement Learning, Deep Q-Network, Representational Similarity Analysis, State-Space Representation, Human Cognition, Atari Games

## 1 Introduction

Deep Reinforcement Learning (DRL) has achieved remarkable success in solving complex decision-making problems directly from visual observations. A prominent example is the Deep Q-Network (DQN), which demonstrated human-level performance in several Atari games using only raw pixel inputs [1]. Despite these achievements, the internal representations learned by such models remain difficult to interpret. Understanding how reinforcement learning agents encode information about their environment is therefore an important challenge for both artificial intelligence and cognitive science.

In reinforcement learning and neuroscience, behavior can be described in terms of state-space representations, where a task is represented as a set of states connected through actions and associated outcomes. Studying whether artificial agents develop representational structures related to these task properties can provide insights into how artificial systems organize information during learning.

In this context, a game state is defined as a short temporal snapshot of the game, represented as a stack of four consecutive frames, capturing the visual information available to the agent at a given moment during gameplay.

A key question arising from this perspective is not only how representations are formed in DQNs, but also how these representations relate to those used by humans when interpreting the same environments. This comparison can help identify whether similar task structures emerge across biological and artificial systems, and to what degree reinforcement learning agents develop representational organizations that resemble those observed in human cognition.

## 2 Research Objectives and Expected Contributions

This project investigates the degree of alignment between human perceptual representations of Atari game states and the internal representations learned by DQNs across different levels of its processing hierarchy. To this end, two specific objectives are pursued:

- **Characterize internal representations of the DQN.** The project analyzes representations produced by different network layers when processing game states, determining whether the network develops a visual-to-abstract gradient in which early layers remain close to raw pixel similarity while deeper layers progressively encode more abstract properties of the game states.
- **Compare human and artificial representations of task states.** Human perceptual representations are

• Contact E-mail: Oscar.Arrocha@autonoma.cat
• Supervised by: Daniel Pacheco (Basic Psychology Area)
• Academic Year 2025/26

measured through two online behavioral experiments using a triplet odd-one-out similarity judgment task on gameplay clips. These judgments are used to construct representational similarity matrices, which are then compared with DQN layer activations using Representational Similarity Analysis (RSA) [2]. The central hypothesis is that human representational structures will align most strongly with the fully connected layer of the DQN, as the most abstract level of its hierarchy.

The expected contributions are threefold: a systematic analysis of how perceptual and abstract information is organized across DQN layers; empirical evidence on the degree of alignment between human perceptual representations and the hierarchical structure learned by the agent; and a contribution to the broader effort to bridge machine learning models and human cognition through common representational frameworks.

# 3 State of the Art

A closely related study is presented in *Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments* by Cross et al. [3]. In that work, participants played Atari games while their brain activity was recorded using fMRI, and a DQN trained on the same games was used as a computational model of behavior and neural activity.

Using RSA, the authors showed that both the human brain and the DQN transform raw pixel inputs into more compact representations that capture task-relevant features, such as the spatial positions of objects important for decision-making. Moreover, activity in deeper layers of the DQN was found to predict neural responses in distributed sensorimotor brain regions involved in action selection.

Similar to that approach, the present project also analyzes internal representations of DQNs using RSA. However, instead of relying on neural recordings, the proposed study focuses on human behavioral similarity judgments. By comparing human-derived representational structures with those extracted from different layers of the DQNs, the project aims to investigate whether human perception of gameplay situations aligns with the abstract representations learned by deep reinforcement learning agents.

# 4 Methodology

## 4.1 Experimental Environment

The experiments in this project are conducted using the Arcade Learning Environment (ALE), a widely used benchmark platform for reinforcement learning research. Three Atari games are used: *Pong*, *Ms. Pac-Man*, and *Space Invaders*. These games were selected because they involve different perceptual and decision-making demands.

Pong requires motion prediction and continuous control, Ms. Pac-Man involves navigation and planning in a visually complex environment, and Space Invaders requires rapid responses to dynamically appearing threats. Using games with different task structures allows the study of how representations learned by the agent vary across environments with different perceptual and behavioral demands.

## 4.2 Deep Reinforcement Learning Agent

The artificial agent used in this study is based on the Deep Q-Network (DQN) architecture introduced by Mnih et al. [1]. For each game environment, five independent agents are trained using the same configuration but different random seeds, following a shared training configuration derived from prior work, with minor adaptations described below. One agent, trained with seed 42 and referred to throughout as the reference seed, was used for all stimulus selection, triplet generation, and difficulty scoring; the remaining four served exclusively for the reliability analyses in Experiments 11–13.

### 4.2.1 State Representation and Preprocessing

The input to the network consists of a stack of four consecutive pre-processed frames. Each frame is constructed from raw Atari observations using the standard preprocessing pipeline for DQN agents.

First, RGB frames are converted to grayscale and downsampled to $84 \times 84$ pixels, with pixel values normalized to the range $[0, 1]$. Following the original DQN procedure, the environment applies frame skipping, where each action is repeated for 4 consecutive game frames. To reduce aliasing effects, the maximum pixel value is taken over the last two skipped frames before downsampling.

Frame stacking is then performed using overlapping windows, such that each state consists of four consecutive processed frames. This representation provides short-term temporal information required for action selection in dynamic environments.

### 4.2.2 Network Architecture

The convolutional feature extractor follows the architecture introduced in the original DQN work [1]. It consists of three convolutional layers followed by a fully connected layer. The first convolutional layer applies 32 filters of size $8 \times 8$ with stride 4, the second layer applies 64 filters of size $4 \times 4$ with stride 2, and the third layer applies 64 filters of size $3 \times 3$ with stride 1. Each convolutional layer is followed by a ReLU activation function. The convolutional backbone is followed by a fully connected layer of 512 units.

### 4.2.3 Training Configuration

The training configuration follows the setup introduced in [3], which uses a DQN agent to study representational structure in high-dimensional reinforcement learning environments. This configuration serves as the baseline for all experiments in this work, and is kept consistent across the three game environments.

Two minor modifications are introduced with respect to the original implementation. First, the replay buffer size is reduced from $1{,}000{,}000$ to $100{,}000$ transitions due to computational constraints. Second, the optimizer is changed from RMSProp to Adam, as implemented in Stable-Baselines3.

## 4.3 Human Similarity Judgment Experiments

To compare human and artificial representations of Atari game states, two online behavioral experiments were designed to collect human similarity judgments of gameplay situations under different coverage and control conditions. The sparse experiment follows an aggregation design, in which a large number of participants each complete a small subset of the available triplets, allowing broad coverage of the stimulus space without placing an excessive burden on any individual participant. The individual experiment instead collects dense judgments from each participant on a fixed, controlled set of trials, enabling the analysis of representational consistency at the level of individual responses. Both experiments shared the same underlying paradigm and instructions, and were implemented using the cognition.run platform (see Appendix A.1 for access links).

In both cases, the behavioral task followed a triplet odd-one-out paradigm. In each trial, participants were presented with three gameplay clips from the same Atari game and were asked to select the clip that was most different from the other two. Participants first received detailed instructions describing the task and the criteria that should guide their judgments. Before each game block, a short gameplay demonstration clip was presented to familiarize participants with the mechanics and objectives of the corresponding game.

Special attention was devoted to the design of the instructions in order to bias participants toward abstract and task-oriented judgments rather than low-level visual similarity. Participants were explicitly instructed to focus on the gameplay situation represented in each clip, including object interactions, movement dynamics, and the relative configuration of game elements. Conversely, they were discouraged from relying exclusively on superficial perceptual features such as score displays, textual information, or small positional differences between objects.

Gameplay clips were extracted from recordings generated by a human player and were designed to approximate the temporal input processed by the DQN while remaining sufficiently interpretable for human participants.

### 4.3.1 Sparse Pilot Experiment

A sparse pilot experiment was conducted to evaluate the feasibility of the similarity judgment paradigm and to identify potential limitations in the experimental pipeline. In this version, clips were manually selected to include diverse gameplay situations for each game.

For each game, ten unique clips were used to generate a pool of 60 triplets. Each clip consisted of 12 raw consecutive RGB frames displayed over approximately 2 seconds. The pilot study included 6 participants in total.

Participant responses from the odd-one-out task were converted into ordinal similarity constraints for the subsequent t-STE embedding procedure (see Section 4.4.3). Each triplet judgment generated two constraints, resulting in a total of 120 constraints per game.

After completing the task, participants answered a short questionnaire regarding prior gaming experience, task clarity, perceived difficulty, and confidence in their responses.

### 4.3.2 Sparse Deployment Experiment

The sparse deployment version incorporates several modifications introduced after the pilot phase in order to improve stimulus selection and triplet generation.

For each game, an optimized 15-clip subset designed to preserve the overall representational structure of the DQN was used. Each clip consisted of 4 consecutive RGB frames, after frame skipping and frame stacking, displayed over 2 seconds.

Triplets were not generated randomly. Instead, they were generated based on distances derived from FC layer representations. For each triplet, a difficulty score was computed based on the relative distance differences between the selected clips, given by the FC layer RDM (see Appendix A.3 for difficulty score computation). This score allowed triplets to be divided into three categories corresponding to easy, medium, and hard similarity judgments.

Each participant received a balanced distribution of triplet difficulties for every game, completing six triplets per game consisting of two easy, two medium, and two hard trials. Triplets were sampled randomly from predefined difficulty pools, and both the order of game blocks and the assignment of clips within triplets were randomized independently for each participant.

Using 15 clips per game results in a total of $\binom{15}{3} = 455$ possible triplets. Based on the reconstruction analysis from Experiment 8, approximately 40% of the available triplets are sufficient to obtain stable representational reconstructions, corresponding to roughly 182 triplets per game. Since each participant completes 6 triplets per game, a minimum of approximately 31 participants is required to reach this level of coverage for each environment. Data collection yielded 39 participants who completed the full experiment, plus partial responses equivalent to approximately 2 additional participants, for an effective total of roughly 41 participants per game.

### 4.3.3 Individual Experiment

The individual experiment was restricted to *Pong* and consisted of 60 triplets drawn as a subset of those presented in the sparse experiment, ensuring that all triplets had already been validated in the broader stimulus pool.

Triplets were stratified by difficulty into three equal groups of 20, corresponding to easy, medium, and hard trials, following the same difficulty scoring described in Appendix A.3. Unlike the sparse experiment, all participants completed the same 60 triplets in the same order. The presentation order was randomized to avoid difficulty-block effects, (easy, medium, and hard trials were interleaved rather than presented in separate blocks) and the assignment of clips to the three positions within each triplet was also randomized. Crucially, both randomizations were shared across participants, such that every participant saw the same triplets, in the same order, with the same clip arrangement.

## 4.4 Representational Similarity Analysis

Representational Similarity Analysis (RSA), introduced by Kriegeskorte et al. [2], is a framework used to compare representations across different systems by analyzing the similarity structure of their responses to the same set of stimuli.

Instead of comparing raw activity patterns directly, RSA compares the pairwise relationships between stimuli in representational spaces. This allows representations derived from artificial neural networks and behavioral data to be analyzed within a common framework.

#### 4.4.1 Representational Dissimilarity Matrices

The central component of RSA is the Representational Dissimilarity Matrix (RDM). An RDM is a square matrix that describes the pairwise dissimilarity between representations of different stimuli. Dissimilarity is computed using a distance metric. The resulting RDM has size $N \times N$, where $N$ is the number of stimuli, and is symmetric with zeros on the diagonal.

In this project, each stimulus corresponds to a game state represented by a stack of four consecutive frames corresponding to the input representation used by the Deep Q-Network.

#### 4.4.2 DQN Representation RDMs

To analyze the representations learned by the Deep Q-Network, RDMs are computed from the activations of different layers of the network.

For each stimulus state $s_i$, the corresponding frame stack is passed through the network and the activation vector of a given layer is extracted. To reduce dimensionality, activation vectors are projected onto their first 100 principal components using Principal Component Analysis (PCA), following the configuration used in previous work [3]. For a layer representation vector $h_i$, the dissimilarity between two states is computed as

$$d_{ij}^{DQN} = 1 - corr(h_i, h_j) \tag{1}$$

where $corr$ denotes the Pearson correlation coefficient. Separate RDMs are computed for each of the four network layers: the three convolutional layers and the fully connected layer, producing layer-specific RDMs that describe how the network organizes game states at different levels of abstraction.

#### 4.4.3 Triplet Embedding with t-STE

Participant responses collected from the sparse experiment are converted into ordinal triplet constraints of the form $(i, j, k)$, indicating that clip $i$ is perceived as more similar to clip $j$ than to clip $k$. These constraints are used to learn a low-dimensional embedding using Stochastic Triplet Embedding (t-STE) [4].

t-STE is a dimensionality reduction method designed to reconstruct geometric representations from relative similarity judgments. The method optimizes an embedding in which the probability of satisfying the observed triplet constraints is maximized. In this work, a separate t-STE embedding is learned for each game using the collected participant triplets.

The resulting embedding provides a continuous geometric representation of the perceived similarities between gameplay clips.

#### 4.4.4 Behavioral Representation RDM

The t-STE embeddings obtained from participant similarity judgments are used to construct behavioral representational dissimilarity matrices (RDMs). For each game, pairwise Euclidean distances are computed between all clip embeddings, resulting in a behavioral RDM that captures the perceived similarity structure of the gameplay states.

These behavioral RDMs are compared with layer-wise RDMs derived from DQN activations using Representational Similarity Analysis (RSA), enabling the identification of network layers whose representations best align with human similarity judgments.

#### 4.4.5 Theoretical Model RDMs

In addition to the DQN layer representations, several theoretical models are constructed to test specific hypotheses about the type of information encoded at different levels of the network and to serve as interpretable reference points in the RSA analysis.

**Pixel Model** The pixel model captures visual similarity between states based on raw low-level image features. Each state is preprocessed following the standard DQN pipeline (grayscale, $84 \times 84$ resolution, normalized to $[0, 1]$) and represented as a flattened pixel vector $v_i \in R^{28224}$. Dissimilarity between two states is computed using correlation distance:

$$d_{ij}^{pixel} = 1 - corr(v_i, v_j) \tag{2}$$

**Pixel PCA Model** The pixel PCA model captures the same low-level visual information as the pixel model but in a reduced representation to its first 100 principal components using PCA. Dissimilarity is then computed using correlation distance over the resulting 100-dimensional representations.

**Handcrafted Feature Model** The handcrafted feature (HCF) model is constructed for Pong only and captures similarity between states based on explicit game-relevant object properties. For each state, a six-dimensional feature vector is extracted from the last two frames of the clip, comprising the ball position ($x$, $y$), ball velocity ($v_x$, $v_y$), and the vertical positions of the left and right paddles. Dissimilarity between two states is computed as the Euclidean distance between their feature vectors:

$$d_{ij}^{HCF} = \|f_i - f_j\|_2 \tag{3}$$

This model provides an interpretable geometric baseline grounded in the actual task structure of Pong, allowing the assessment of whether DQN or human representations reflect the spatial configuration of game objects.

#### 4.4.6 Comparing Representations

RSA is used to compare the representational structures obtained from the different sources.

The similarity between two RDMs is measured by computing the Spearman correlation between the vectorized upper triangular elements of the matrices:

$$RSA(R_1, R_2) = \rho(vec(R_1), vec(R_2)) \quad (4)$$

where $\rho$ denotes the Spearman correlation coefficient.

This procedure allows the identification of which layers of the Deep Q-Network produce representational structures that are most similar to human behavioral judgments or theoretical models.

# 5 Experimental Design

## 5.1 Pilot Experiment

Three experiments were conducted using the pilot sparse task. The first evaluated whether participant similarity judgments produced a consistent low-dimensional representational structure across participants, using t-STE embeddings assessed via 6-fold cross-validation. The second assessed the usability and perceived difficulty of the task through a post-experiment questionnaire. The third performed an initial Human-DQN RSA analysis to investigate whether DQN representations were related to the observed human similarity structure. Full details and results of these experiments are reported in Appendix A.4.

## 5.2 DQN Representation Pipeline Refinement

The preliminary RSA analysis motivated two refinements to the DQN preprocessing pipeline. The representation extraction procedure was updated to faithfully reproduce the full DQN frame-skipping and frame-stacking pipeline, and PCA-based dimensionality reduction was applied to activation vectors prior to RDM computation, following Cross et al. [3]. Full technical details are provided in Appendix A.5 (Experiment 4).

## 5.3 Improvement Experiments

To better understand the source of this variability in RSA results and improve the robustness of the behavioral experiment, a new set of experiments was conducted focusing on stimulus representativeness, clip subset selection, and triplet reconstruction stability.

These experiments aimed to evaluate whether the gameplay clips used in the pilot experiment adequately captured the global representational structure of the DQN, determine the amount of data required to obtain stable representational estimates, and optimize the final experimental design for the deployment version of the study.

### 5.3.1 Experiment 5: Clip Representativeness Experiment

This experiment evaluated whether the manually selected clips used in the pilot experiment were sufficient to reproduce the global representational structure learned by the DQN agents.

For each game, a larger dataset composed of 909 gameplay clips was extracted from human gameplay recordings. Consecutive clips were separated by 50 frames in order to reduce temporal redundancy while maintaining broad coverage of gameplay situations.

Using this dataset, DQN RDMs were computed independently for each network layer following the refined preprocessing pipeline described previously. Layer-wise RSA matrices were then generated for each game.

The RSA structures obtained from the pilot clip subsets were subsequently compared against the RSA structures computed from the full 909-clip dataset. Similarity between both representational organizations was quantified using Spearman correlation between the upper triangular values of the corresponding RSA matrices.

### 5.3.2 Experiment 6: Optimal Number of Clips Experiment

After observing that the initial pilot clips did not fully reproduce the global representational structure of the DQN, an additional experiment was conducted to evaluate how the number of clips affected the stability of the resulting RSA organization.

For each game, random subsets containing different numbers of clips were sampled from the 909-clip dataset. Subset sizes of 10, 12, 15, 20, 25, 30, and 50 clips were evaluated. For each subset size, 100 random subsets were generated independently.

For every subset, DQN RDMs and layer-wise RSA matrices were computed following the refined preprocessing pipeline. The resulting RSA structures were then compared against the RSA matrix obtained from the complete 909-clip dataset using Spearman correlation.

The objective of this experiment was to determine the minimum number of clips required to obtain a stable approximation of the global representational geometry while maintaining a manageable number of triplets needed to be collected.

### 5.3.3 Experiment 7: Optimized Clip Subset Selection

This experiment investigated whether a reduced subset of gameplay clips could be selected to preserve the representational geometry obtained from the large-scale RSA analysis.

For each game, a new pool of 25 candidate clips was manually selected to provide broader coverage of gameplay situations. From this pool, subsets containing 15 clips were evaluated according to their ability to reproduce the RSA structure obtained from the 909-clip reference dataset.

For every candidate subset, DQN RDMs and layer-wise RSA matrices were computed following the refined preprocessing pipeline. Similarity between subset RSA matrices and the reference RSA structure was quantified using Spearman correlation between the upper triangular values of the corresponding RSA matrices.

An iterative subset optimization procedure based on simulated annealing was then applied to progressively improve the representativeness of the selected clip subsets by repeatedly replacing candidate clips and reevaluating the resulting RSA similarity.

### 5.3.4 Experiment 8: Triplet Reconstruction Analysis

After selecting representative gameplay clip subsets, an additional experiment was conducted to estimate the number

of triplet constraints required to reliably reconstruct DQN representational geometries using t-STE embeddings.

For each game, triplet constraints were generated directly from the DQN fully connected layer RDM. Different percentages of the available triplets were then randomly sampled, ranging from 10% to 100% of the total constraints.

For each sampled subset, a t-STE embedding was computed and converted back into an RDM. Reconstruction quality was measured using Spearman correlation between the reconstructed RDM and the original DQN RDM. The procedure was repeated 100 times for each condition in order to estimate both reconstruction performance and variability.

The objective of this experiment was to determine the amount of behavioral data required to obtain stable human representational dissimilarity matrices for the final deployment experiment.

## 5.4 Human Response Consistency

Before comparing human and DQN representational structures, it is necessary to establish that the collected behavioral data is reliable enough to support such a comparison. If human responses were largely inconsistent across participants, the resulting RDM would reflect noise rather than a shared representation, making any correlation with DQN representations uninterpretable. Two experiments were therefore conducted to verify that participants agree with each other in their similarity judgments, and that the two behavioral paradigms, the individual and sparse experiments, converge on the same representational structure over the same set of triplets. Both analyses are restricted to Pong.

### 5.4.1 Experiment 9: Individual Experiment Internal Consistency

This experiment evaluated the degree to which participants in the individual experiment agreed with each other in their similarity judgments. Since all participants completed the same 60 triplets, direct response-level comparisons across participants were possible.

For each participant and difficulty level, consistency was defined as the proportion of triplets on which the participant's response matched the plurality vote of the remaining participants. The plurality vote was computed using a leave-one-out procedure: for each triplet, the most common response among all participants except the one being evaluated was taken as the reference, and the participant's response was scored as consistent if it matched this reference. This procedure avoids circularity by ensuring that the reference is never influenced by the participant being evaluated.

Group-level consistency per difficulty was summarized as the mean plurality rate across triplets, where the plurality rate of a triplet is the fraction of participants who selected the most common response. A higher plurality rate indicates stronger agreement among participants and a more reliable collective judgment. Statistical significance was assessed using a one-sample $t$-test against the chance level of $1/3$.

### 5.4.2 Experiment 10: Sparse–Individual Agreement

This experiment assessed whether the representational structure captured by the sparse experiment is consistent with that captured by the individual experiment, despite the two paradigms differing in participant coverage, triplet assignment, and data density.

The comparison was restricted to the 60 triplets shared between both experiments. For each of these triplets, a majority vote was computed independently from the sparse and individual experiment responses. Agreement between the two experiments was then quantified as the proportion of triplets for which the majority vote was identical across paradigms, computed separately for each difficulty level. Statistical significance was assessed using a one-sided binomial test against the chance level of $1/3$.

## 5.5 DQN Reliability and Noise Ceiling

Having established that human similarity judgments are internally consistent, a parallel set of analyses was conducted on the model side. Before any human–DQN comparison can be interpreted, it is necessary to verify that DQN representations are themselves stable across independent training runs, and to establish an upper bound for the correlations that the sparse triplet protocol can recover. Three experiments were therefore conducted in sequence: assessing cross-seed representational agreement, investigating where and why seeds diverge as a function of triplet difficulty, and estimating the noise ceiling for the subsequent human–DQN RSA analysis.

### 5.5.1 Experiment 11: Cross-Seed RDM Consistency

To assess whether a single DQN seed can serve as a representative reference model, the pairwise RSA was computed between five independently trained agents, all with the same training configuration but different seed. For each game and each network layer, RDMs were computed from the activations of each seed following the preprocessing and PCA reduction pipeline described in Section 4.4.2. Pairwise Spearman RSA was then performed between all $\binom{5}{2} = 10$ seed pairs, yielding a $5 \times 5$ cross-seed RSA matrix per game and layer. The mean and standard deviation of the off-diagonal entries were reported as a reliability index for each condition.

### 5.5.2 Experiment 12: Triplet Difficulty and Cross-Seed Agreement

To investigate whether the FC layer cross-seed divergence observed in Experiment 11 is concentrated in structurally ambiguous triplets, the difficulty-bucketed triplets derived from reference seed's FC layer RDM were used (see Appendix A.3 for the full bucketing procedure). For each non-reference seed, the predicted odd clip was recorded for every triplet in the hard, medium, and easy buckets, and agreement with reference seed was computed as the mean fraction of seeds selecting the same odd clip per bucket.

### 5.5.3 Experiment 13: DQN Noise Ceiling

To provide an upper bound for the human–DQN RSA analysis, a noise ceiling was estimated using the sparse triplets

presented to human participants. For each non-reference seed, the synthetic triplet responses recorded in Experiment 12 were restricted to the sparse subset and used to reconstruct a t-STE embedding following the same procedure applied to human data: 100 runs with random initializations, retaining the run whose RDM best correlated with the reference seed's FC layer RDM. The Spearman correlation between the retained t-STE RDM and the reference seed's RDM was recorded per seed, and the noise ceiling was defined as the mean across the four non-reference seeds.

## 5.6 DQN and Human Agreement by Triplet Difficulty

The preceding experiments established that human observers are internally consistent in their similarity judgments and that DQN representations are stable across independent training runs. Both sources of representation have therefore been validated in isolation. This section investigates whether DQN-derived triplet difficulty predicts human behavioral difficulty: specifically, whether human agreement with the ground-truth odd clip follows the same gradient across easy, medium, and hard triplets as cross-seed DQN agreement.

### 5.6.1 Experiment 14: Human Agreement with Ground Truth by Triplet Difficulty

For each game, the majority-vote response of the sparse experiment was compared against the ground-truth odd clip (the reference seed responses) across the three difficulty buckets. Agreement was quantified using the tie-aware procedure introduced in Experiment 10, yielding observed, minimum, and maximum agreement values per bucket. Statistical significance was assessed per bucket using a one-sided binomial test against chance $(1/3)$.

## 5.7 DQN Representation Analysis

To investigate the representational structure learned by the DQN, a series of RSA comparisons were conducted using the theoretical models defined in Section 4.4.5. All analyses are based on a single RSA matrix computed simultaneously across all RDMs, including DQN layer representations, theoretical models, and the human sparse RDM. This allows a unified comparison of how different sources of information relate to each other and to the internal organization of the network.

### 5.7.1 Visual-to-Abstract Gradient across DQN Layers

This experiment investigated whether the DQN develops a progression from perceptual to abstract representations across its layers. The hypothesis is that early convolutional layers encode primarily low-level visual features, while the fully connected layer captures more abstract, task-relevant structure.

To test this, the pixel RDM was used as a reference for low-level visual similarity, and the HCF RDM (Pong only) was used as a reference for abstract, object-level structure. RSA correlations between each of these models and each DQN layer were compared across the network hierarchy. A visual-to-abstract gradient would manifest as a decreasing correlation with the pixel RDM and an increasing correlation with the HCF RDM from early to late layers.

### 5.7.2 Human-DQN Alignment

This experiment assessed the degree to which human similarity judgments align with the representational structure of the DQN and with the theoretical models. The human sparse RDM was compared against all DQN layer RDMs and all theoretical model RDMs using RSA. This analysis allows the identification of which layer of the network best captures human perception of gameplay situations, and whether human representations are more consistent with low-level visual similarity or with higher-level, task-relevant structure.

# 6 Results

## 6.1 Pilot Experiment and Pipeline Refinement

The pilot study yielded consistent low-dimensional embeddings across participants, with two dimensions providing the best generalization performance in all three games (see Appendix A.4 for full cross-validation results and questionnaire analysis). Post-experiment feedback confirmed that the task was interpretable, though two participants noted the experiment felt somewhat long, motivating a reduction in the number of triplets per participant for the deployment version.

A preliminary RSA analysis on the pilot data revealed interpretable patterns (notably a tendency toward higher human–DQN correlation at the fully connected layer in Pong) while also indicating that more robust conclusions would require a more representative stimulus set and better calibrated experimental parameters. Following pipeline refinement, the recomputed RSA results showed partial alignment between human similarity judgments and DQN representations, most notably at the fully connected layer in Pong ($\rho = 0.297$) and across convolutional layers in Space Invaders ($\rho = 0.464$–$0.674$), but with inconsistent patterns across games. These observations motivated the systematic improvement experiments described in the following subsection (see Appendix A.5.4 for the full preliminary RSA results).

## 6.2 Improvement Experiments

### 6.2.1 Experiment 5: Clip Representativeness

Table 1 shows the correlation between the RSA structures obtained using the pilot clips and those computed from the larger 909-clip dataset.

| Game | RSA Correlation |
|---|---|
| Pong | 0.88 |
| Ms. Pac-Man | 0.60 |
| Space Invaders | 0.37 |

TABLE 1: Correlation between pilot RSA structures and large-scale RSA structures.

The results show that the selected clips for Pong reproduce the overall RSA organization of the DQN with high similarity. In contrast, lower correlations were obtained for Ms. Pac-Man and especially for Space Invaders, suggesting that the manually selected clips were not sufficiently representative of the full DQN representation space for these games.

These findings indicate that clip selection substantially influences the resulting RSA structure and suggest that future experiments should incorporate improved clip sampling strategies capable of better capturing the diversity of gameplay situations and the global representational organization of the DQN.

#### 6.2.2 Experiment 6: Optimal Number of Clips

RSA stability increased with the number of clips but showed diminishing returns beyond 15, which was therefore selected as the stimulus set size to balance representational coverage against experimental feasibility (see Appendix A.6.1 for the full scaling results).

#### 6.2.3 Experiment 7: Optimized Clip Subset Selection

The experiment produced optimized subsets of 15 clips for all three games. In all cases, the resulting RSA matrices achieved a Spearman correlation of 1.0 with the RSA computed from the full 909-clip dataset. These results indicate that carefully selected subsets can preserve the overall representational structure of the DQN while considerably reducing the number of stimuli required for human experiments.

The optimized 15-clip subsets were therefore adopted for the final deployment version of the behavioral task.

#### 6.2.4 Experiment 8: Triplet Reconstruction Analysis

Reconstruction quality stabilized after approximately 40% of available triplet constraints, with full-constraint correlations of $0.69$, $0.87$, and $0.90$ for Pong, Ms. Pac-Man, and Space Invaders respectively, confirming that reliable approximations of the DQN representational geometry are achievable under the sparse behavioral protocol (see Appendix A.6.2 for full results).

### 6.3 Human Response Consistency

#### 6.3.1 Experiment 9: Individual Experiment Internal Consistency

Table 2 reports the mean per-participant consistency in the individual experiment across difficulty levels, defined as the proportion of triplets on which each participant's response matched the leave-one-out plurality vote of the remaining participants.

Consistency exceeded chance across all difficulty levels ($p < 0.001$ in all cases), and every participant individually exceeded chance in all three conditions. Easy triplets produced the highest and most uniform agreement (mean $= 0.85$, range $[0.75, 1.00]$). Medium and hard triplets showed lower but still substantial consistency (means of $0.62$ and $0.66$ respectively), with greater individual variation. Hard triplets produced slightly higher consistency than medium ones ($0.661$ vs $0.622$), though the difference was not tested for significance given the small sample size.

| Difficulty | Mean | Std | $p$ |
|---|---|---|---|
| Easy | 0.850 | 0.103 | $<0.001$ |
| Medium | 0.622 | 0.133 | $<0.001$ |
| Hard | 0.661 | 0.132 | $<0.001$ |

TABLE 2: Mean per-participant consistency (proportion of responses matching the LOO plurality vote) for each difficulty level in the individual experiment. $p$-values are from one-sample $t$-tests against chance ($1/3$). $n = 9$ participants.

#### 6.3.2 Experiment 10: Sparse–Individual Agreement

To assess whether the shared structure shown in Experiment 9 generalizes across paradigms, Table 3 reports the agreement between the majority votes of the sparse and individual experiments on the 60 shared Pong triplets. Because many sparse triplets received only two responses, tied votes, where both responses selected different clips, cannot be resolved by majority. The table therefore reports the observed agreement under arbitrary tie-breaking, as well as the minimum and maximum possible agreement across all valid tie-breaking choices, together with the number of tied triplets per condition.

| Diffic. | Ties | Min | Obs. | Max | $p_{\min}$ | $p_{\max}$ |
|---|---|---|---|---|---|---|
| Easy | 3 | 0.65 | 0.80 | 0.80 | 0.004 | $<0.001$ |
| Medium | 4 | 0.45 | 0.50 | 0.65 | 0.191 | 0.004 |
| Hard | 4 | 0.45 | 0.50 | 0.55 | 0.191 | 0.038 |

TABLE 3: Agreement between sparse and individual experiment majority votes on the 60 shared Pong triplets, by difficulty. Ties: number of sparse triplets where two responses selected different clips. Min and Max: lowest and highest possible agreement under any tie-breaking choice. $p$-values correspond to one-sided binomial tests against chance ($1/3$) for the min and max agreement values. $n = 20$ triplets per condition.

For easy triplets, agreement was significant regardless of how ties are resolved (min $= 0.65$, $p = 0.004$; max $= 0.80$, $p < 0.001$), indicating that the two paradigms converge on the same majority response for discriminable stimuli. For medium and hard triplets, the result depends on the tie-breaking outcome: under the most favorable resolution, agreement reaches $0.65$ ($p = 0.004$) and $0.55$ ($p = 0.038$) respectively, while under the least favorable resolution neither condition differs significantly from chance. The observed values of $0.50$ in both cases fall between these bounds and do not reach significance. The ambiguity introduced by tied sparse responses therefore prevents a definitive conclusion about cross-paradigm agreement for harder triplets.

Together, these results indicate that participants share a consistent representational structure for Pong, as evidenced by the internal consistency of the individual experiment across all difficulty levels. The agreement between sparse and individual majority votes for easy triplets further supports that the sparse experiment captures this shared rep-

resentation. For medium and hard triplets the evidence is inconclusive due to the limited number of sparse responses per triplet in those conditions. The sparse experiment RDM is therefore used in subsequent analyses as an estimate of the shared human representational structure, acknowledging that the cross-paradigm validation was only confirmed for the most discriminable stimuli.

## 6.4 DQN Reliability and Noise Ceiling

### 6.4.1 Experiment 11: Cross-Seed RDM Consistency

Table 4 reports the mean and standard deviation of the pairwise cross-seed RSA correlations for each game and layer.

| Game | Layer | Mean $\rho$ | Std |
|---|---|---|---|
| Pong | Conv. 1 | 0.798 | 0.038 |
| | Conv. 2 | 0.833 | 0.033 |
| | Conv. 3 | 0.799 | 0.021 |
| | FC | 0.391 | 0.020 |
| Ms. Pac-Man | Conv. 1 | 0.955 | 0.010 |
| | Conv. 2 | 0.889 | 0.017 |
| | Conv. 3 | 0.898 | 0.023 |
| | FC | 0.729 | 0.052 |
| Space Invaders | Conv. 1 | 0.945 | 0.009 |
| | Conv. 2 | 0.845 | 0.012 |
| | Conv. 3 | 0.774 | 0.010 |
| | FC | 0.570 | 0.009 |

TABLE 4: Cross-seed RSA reliability. Mean and standard deviation of Spearman correlations across all 10 pairwise seed comparisons, per game and layer.

Cross-seed RSA was consistently high in the convolutional layers across all three games, with means ranging from $0.77$ to $0.96$, indicating that the low-level visual representations learned by the DQN are largely determined by the shared architecture and stimulus statistics rather than by the specific training trajectory. The fully connected layer, however, shows a pronounced drop in cross-seed agreement across all games: from the convolutional baseline down to $0.391 \pm 0.020$ for Pong, $0.729 \pm 0.052$ for Ms. Pac-Man, and $0.570 \pm 0.009$ for Space Invaders. This is noteworthy given that all five seeds achieve comparable in-game performance despite this representational divergence at the FC layer (see Appendix A.8).

One possible interpretation is that the abstract FC layer representation is underdetermined by the task: multiple geometrically distinct organizations may yield equally high reward, allowing different seeds to diverge. The drop is most severe for Pong, consistent with its simpler visual structure leaving the FC layer geometry particularly underspecified. Whether this divergence is uniformly distributed across all triplets or concentrated in structurally ambiguous ones is the question addressed by Experiment 12.

### 6.4.2 Experiment 12: Triplet Difficulty and Cross-Seed Agreement

Table 5 reports cross-seed agreement with the reference seed broken down by difficulty bucket and game.

Agreement follows the expected gradient across all three games, with easy triplets yielding the highest cross-seed agreement and hard triplets the lowest. This mirrors the pattern observed in human participants in Experiment 9, suggesting that the difficulty structure reflects genuine geometric properties of the representational space shared across both agents and observers.

| Game | Difficulty | Mean agreement | Std |
|---|---|---|---|
| Pong | Easy | 0.914 | 0.192 |
| | Medium | 0.503 | 0.305 |
| | Hard | 0.465 | 0.306 |
| Ms. Pac-Man | Easy | 0.753 | 0.332 |
| | Medium | 0.630 | 0.336 |
| | Hard | 0.469 | 0.411 |
| Space Invaders | Easy | 0.952 | 0.115 |
| | Medium | 0.866 | 0.232 |
| | Hard | 0.653 | 0.357 |

TABLE 5: Mean fraction of non-reference seeds agreeing with the reference seed per difficulty bucket and game. Std denotes standard deviation across triplets within each bucket.

The effect is most pronounced in Pong, where agreement collapses to near chance for both hard and medium buckets ($0.465$ and $0.503$ respectively), while remaining high for easy triplets ($0.914$). This is directly linked to the difficulty score ranges: Pong's hard triplets have scores in the range $[0.017, 0.040]$, substantially lower than those of Pac-Man $[0.048, 0.103]$ and Space Invaders $[0.048, 0.103]$, meaning the three clips occupy a nearly equidistant cluster where minor representational differences across seeds suffice to reverse the identity of the odd clip. This explains the low overall FC layer RSA for Pong observed in Experiment 11: the disagreement is not broadly distributed but is driven by a subset of geometrically ambiguous triplets.

For easy triplets, seeds converge reliably across all three games. For hard and medium triplets, the near-chance agreement reflects geometric ambiguity intrinsic to the representational space rather than an arbitrary choice of the reference seed, no seed is more correct than another when the three clips are nearly equidistant. The reference seed is therefore a valid representative of the model class: where the geometry is well-defined, all seeds agree with it, and where they do not, the disagreement is a property of the triplet structure itself. To verify that this pattern is not an artefact of the three-bucket discretisation, the analysis was repeated on the continuous difficulty score axis using ten equal-width bins across all 455 triplets. The monotonic gradient is preserved in full for all three games: for Pong, agreement rises from $0.469$ at $\delta \approx 0.05$ to $1.000$ at $\delta \geq 0.55$; for Ms. Pac-Man from $0.444$ to $0.896$; and for Space Invaders from $0.546$ to $1.000$ (see Appendix A.9 for the full binned table).

### 6.4.3 Experiment 13: DQN Noise Ceiling

Table 6 reports the per-seed correlations and resulting noise ceiling for each game.

The noise ceiling varies substantially across games: Space Invaders reaches $0.806 \pm 0.026$, Ms. Pac-Man $0.554 \pm 0.068$, and Pong $0.423 \pm 0.062$. The low ceiling for Pong is consistent with the findings of Experiments 11 and 12:

| Game | Noise ceiling | Std |
|---|---|---|
| Pong | 0.423 | 0.062 |
| Ms. Pac-Man | 0.554 | 0.068 |
| Space Invaders | 0.806 | 0.026 |

TABLE 6: Noise ceiling (mean $\pm$ std Spearman correlation across four non-reference seeds) per game.

the near-equidistant triplet structure and the resulting cross-seed divergence on hard and medium triplets limit how much of the reference seed's RDM can be recovered from sparse triplet data, even by a fully trained DQN. Any human–DQN correlation will therefore be interpreted relative to these ceilings rather than as an absolute value.

## 6.5 DQN and Human Agreement by Triplet Difficulty

### 6.5.1 Experiment 14: Human Agreement with Ground Truth by Triplet Difficulty

Table 7 reports the results across all three games and difficulty buckets.

| Game | Bucket | Obs. | Min | Max | $p_{\text{obs}}$ |
|---|---|---|---|---|---|
| Pong | Easy | 0.776 | 0.714 | 0.857 | <0.001 |
| | Medium | 0.391 | 0.326 | 0.478 | 0.246 |
| | Hard | 0.438 | 0.354 | 0.500 | 0.086 |
| Ms. Pac-Man | Easy | 0.380 | 0.320 | 0.460 | 0.287 |
| | Medium | 0.440 | 0.380 | 0.480 | 0.076 |
| | Hard | 0.357 | 0.304 | 0.429 | 0.401 |
| Space Invaders | Easy | 0.583 | 0.500 | 0.667 | <0.001 |
| | Medium | 0.686 | 0.647 | 0.784 | <0.001 |
| | Hard | 0.577 | 0.519 | 0.596 | <0.001 |

TABLE 7: Human majority-vote agreement with the ground-truth odd clip per difficulty bucket and game. Obs.: observed agreement under arbitrary tie-breaking. Min/Max: bounds across all valid tie-breaking choices. $p_{\text{obs}}$: one-sided binomial test against chance $(1/3)$ on the observed value. $n$: number of sparse triplets matched per bucket.

The results differ substantially across games, reflecting the heterogeneity in human representational consistency already noted in earlier experiments.

*Pong* shows the clearest parallel with the DQN difficulty structure. Easy triplets produced a strong and unambiguous agreement of $0.776$ ($p < 0.001$), well above chance and consistent with the high cross-paradigm agreement observed for easy triplets in Experiment 10 (min $= 0.65$, $p = 0.004$). Medium and hard triplets both fall to near-chance levels ($0.391$ and $0.438$ respectively), with neither reaching significance on the observed value ($p = 0.246$ and $p = 0.086$). This mirrors the collapse in cross-seed DQN agreement at hard and medium triplets reported in Experiment 12 ($0.465$ and $0.503$ respectively), where near-chance DQN agreement was linked to the extremely compressed difficulty score range of Pong's hard triplets ($[0.017, 0.040]$). The slight non-monotonicity between medium and hard buckets (hard agreement marginally exceeding medium) echoes the same pattern observed in Experiment 9, where hard triplets produced marginally higher internal consistency than medium ones ($0.661$ vs $0.622$), and should not be interpreted as a meaningful reversal given the overlapping confidence bounds and the small difference in observed values ($0.047$). Taken together, Pong provides the strongest evidence that DQN-derived difficulty predicts human behavioral difficulty: both signals are above chance only for easy triplets and collapse together at the harder end of the axis.

*Space Invaders* presents a different and equally informative pattern. Human agreement exceeds chance significantly across *all three* difficulty buckets: easy $0.583$, medium $0.686$, and hard $0.577$ (all $p < 0.001$). Rather than a gradient of deterioration, performance is relatively uniform across difficulty levels, suggesting that human observers maintain consistent judgments regardless of the geometric separation between clips in the DQN space. This is consistent with the high noise ceiling estimated in Experiment 13 ($0.806 \pm 0.026$), which indicated that the sparse triplet protocol can recover a substantial proportion of the reference seed's representational geometry for this game, and with the high reconstruction fidelity reported in Experiment 8 ($\rho = 0.90$). The result implies that for Space Invaders, the DQN difficulty axis does not straightforwardly translate into behavioral difficulty for humans, possibly because the perceptual structure of the game (regular enemy formations and predictable motion) provides additional cues that allow humans to resolve even geometrically ambiguous triplets consistently.

*Ms. Pac-Man* shows no significant above-chance agreement at any difficulty level on the observed value: easy $0.380$ ($p = 0.287$), medium $0.440$ ($p = 0.076$), and hard $0.357$ ($p = 0.401$). This failure to exceed chance is not explained by difficulty alone; it holds even for easy triplets, which produced strong performance in both Pong and Space Invaders. The result is instead consistent with the structural limitations of the Pac-Man data identified across multiple earlier experiments: the relatively low t-STE generalization accuracy in the deployment cross-validation (Section 5.2), the lower clip representativeness correlation in Experiment 5 ($\rho = 0.60$), and the lowest noise ceiling of the three games ($0.554 \pm 0.068$, Experiment 13). Together these observations suggest that human similarity judgments for Ms. Pac-Man are less coherently organized around the DQN representational geometry, and that the absence of a difficulty gradient here reflects a broader misalignment between human and DQN representational axes for this game rather than a failure of the difficulty scoring procedure.

Across games, human–DQN difficulty correspondence is game-dependent: strongest in Pong, absent in Ms. Pac-Man, and inverted in Space Invaders where humans outperform the DQN difficulty prediction, motivating the game-by-game RSA interpretation in Section 6.6.

## 6.6 DQN Representation Analysis

The full RSA matrices for all three games are provided in Appendix A.11.

### 6.6.1 Visual-to-Abstract Gradient across DQN Layers

In *Pong*, the pixel RDM correlates most strongly with Conv. 1 ($\rho = 0.451$) and decreases monotonically toward the FC layer ($\rho = 0.141$), with the pixel PCA model show-

ing a sharper decline from Conv. 1 ($\rho = 0.504$) to near zero at the FC layer ($\rho = -0.077$). The HCF model does not follow the expected complementary pattern: rather than increasing toward the FC layer, it peaks at Conv. 2 ($\rho = 0.418$) and Conv. 3 ($\rho = 0.395$), and drops to $\rho = 0.103$ at the FC layer. This suggests that while the FC layer has moved away from visual representations, it has not reorganized around the explicit geometric task features captured by the HCF model. Given the low cross-seed FC reliability observed for Pong in Experiment 11 ($\rho = 0.391$), a plausible interpretation is that the FC layer geometry in Pong is underspecified by the task: multiple geometrically distinct organizations yield equivalent policy performance, and the particular solution found by the reference seed does not align with the HCF feature space.

In *Ms. Pac-Man*, both pixel models show a clear decreasing trend from Conv. 1 ($\rho = 0.782$ and $\rho = 0.795$ for pixel and pixel PCA respectively) toward the FC layer ($\rho = 0.477$ and $\rho = 0.373$). The gradient is present but shallower than in Pong, with the FC layer retaining a moderate positive correlation with visual similarity. This is consistent with the higher visual complexity of Ms. Pac-Man, where low-level features such as the spatial layout of the maze, pellet positions, and ghost configurations remain partially informative at higher levels of the network.

In *Space Invaders*, the pixel RDM shows a strong decline from Conv. 1 ($\rho = 0.321$) to near zero at the FC layer ($\rho = -0.077$), suggesting a sharp decoupling from raw pixel similarity. The pixel PCA model, however, tells a different story: it correlates strongly with Conv. 1 ($\rho = 0.822$) and decreases more gradually toward the FC layer ($\rho = 0.437$), indicating that the network remains sensitive to the dominant visual structure of the game even at deeper layers. The large discrepancy between the two pixel models (absent in Pong and Pac-Man) reflects the highly regular visual structure of Space Invaders, where a small number of principal components captures most of the frame variance. Raw pixel distances are a poor proxy for this structure, while the PCA projection recovers it.

Taken together, these results provide evidence for a visual-to-abstract gradient across all three games, clearest in Pong and present but shallower in Ms. Pac-Man and Space Invaders due to their greater visual complexity. In Pong, the absence of a corresponding HCF increase at the FC layer suggests the abstract representations learned by the network are not straightforwardly organized around the explicit task-relevant geometric features that define optimal play.

### 6.6.2 Human-DQN Alignment

In *Pong*, human similarity judgments show near-zero or negative correlations with the convolutional layers (Conv. 1: $\rho = -0.090$, Conv. 2: $\rho = 0.058$, Conv. 3: $\rho = 0.213$) and a moderate positive correlation with the FC layer ($\rho = 0.324$). Among all reference models, the strongest alignment is with the HCF model ($\rho = 0.499$), which captures the spatial configuration of the ball and paddles. Human representations in Pong are therefore more consistent with explicit task-relevant geometric features than with any layer of the DQN. The FC layer shows only weak alignment with HCF ($\rho = 0.103$), indicating that the partial human–FC correspondence cannot be attributed to the geometric task features encoded in the HCF model, since the FC layer itself does not organize states according to those features. Given the noise ceiling of $0.423$ estimated in Experiment 13, the observed human–FC correlation of $0.324$ approaches the upper bound recoverable with the sparse triplet protocol, suggesting that the true alignment may be constrained more by the geometric ambiguity of the FC layer than by a genuine absence of shared structure.

In *Ms. Pac-Man*, human correlations are low across all layers and reference models. The highest value is observed at Conv. 1 ($\rho = 0.175$), with a monotonic decrease toward the FC layer ($\rho = 0.005$). The pixel RDM produces the highest overall alignment with humans ($\rho = 0.205$), marginally above the convolutional layers. These results indicate no reliable alignment between human similarity judgments and DQN representations at any level of the network, and are consistent with the structural limitations identified across earlier experiments: the low clip representativeness correlation (Experiment 5, $\rho = 0.60$), the low noise ceiling ($0.554 \pm 0.068$), and the absence of a significant difficulty gradient in Experiment 14. Together these observations suggest that human representational axes for Ms. Pac-Man do not correspond to those captured by the DQN geometry under the current experimental design.

In *Space Invaders*, human similarity judgments show strong correlations across all DQN layers, with values ranging from $\rho = 0.630$ at Conv. 2 to $\rho = 0.690$ at Conv. 1, and $\rho = 0.684$ at the FC layer. Notably, the raw pixel RDM produces near-zero alignment ($\rho = -0.032$), while the pixel PCA model yields a moderate correlation ($\rho = 0.438$). This pattern indicates that human judgments in Space Invaders are not driven by unstructured pixel-level similarity but align with the principal axes of visual variation captured by both the pixel PCA model and the DQN across all its layers. The absence of a layer-specific gradient, with the FC layer performing no better than Conv. 1, suggests that the alignment does not reflect sensitivity to abstract task representations in particular, but rather to a mid-level visual structure that the network encodes consistently throughout its hierarchy. This interpretation is supported by the high noise ceiling for this game ($0.806 \pm 0.026$) and the high reconstruction fidelity reported in Experiment 8 ($\rho = 0.90$), confirming that the strong human–DQN alignment observed here is reliable and not an artefact of the sparse triplet protocol.

Across games, the results reveal a heterogeneous picture. A partial human–DQN alignment exists in Pong, localized to the FC layer and likely bounded by the geometric instability of that layer across training seeds. In Space Invaders, alignment is strong and layer-independent, pointing to shared sensitivity to mid-level visual structure rather than to abstract task representations specifically. In Ms. Pac-Man, no reliable alignment is observed at any level. These contrasts indicate that the degree and nature of human–DQN correspondence is strongly modulated by the perceptual and structural properties of the game environment.

## 7 Conclusions

This project investigated the alignment between human perceptual representations of Atari game states and the internal representations of a Deep Q-Network, using a triplet odd-

one-out paradigm and Representational Similarity Analysis. The results reveal a heterogeneous picture across environments, summarized below.

## 7.1 Summary of Main Findings

Human participants showed consistent, above-chance representational structure for all three games, and the DQN exhibited a visual-to-abstract gradient across its layers in all three environments. However, the nature and degree of human–DQN correspondence differed substantially by game: in Pong, partial alignment was localized to the FC layer and bounded by its geometric instability across seeds; in Space Invaders, strong layer-independent alignment reflected shared sensitivity to mid-level visual structure rather than abstract representations specifically; and in Ms. Pac-Man, no reliable alignment was found at any level, consistent with the structural limitations of the experimental design for that environment.

## 7.2 Limitations and Possible Causes

Several factors may have contributed to the weaker-than-expected RDM alignment, particularly for Ms. Pac-Man and Space Invaders.

A key methodological consideration concerns differences between the DQN training configuration used in this work and the one employed in the original Cross et al. [3] study. While the overall architecture and hyperparameter choices were kept as close as possible to that reference, two modifications were introduced due to computational constraints: the replay buffer was reduced from 1,000,000 to 100,000 transitions, and the optimizer was changed from RMSProp to Adam. These differences may have affected the representational geometry of the trained agents, particularly at the FC layer, potentially producing representations that diverge from those reported in the reference work. The reduced replay buffer in particular limits the diversity of transitions available during training, which could influence the degree to which the agent develops abstract, task-relevant state representations rather than representations dominated by recency or local visual statistics.

A second limitation relates to the nature of the reward structure in Ms. Pac-Man and Space Invaders. In these games, the most critical outcomes (being caught by a ghost, or allowing enemies to reach the lower screen border) are rare and delayed events within each episode. As a result, the reward signal that drives the DQN's representation learning is sparse with respect to those decisive task states. The agent may thus spend most of its training without reinforcing the abstract situational distinctions that humans consider most salient. This could explain why, for both games, the FC layer retains a moderate correlation with visual models even at depth: the network has not been sufficiently driven to reorganize its representations around abstract task-relevant structure, and low-level visual features continue to dominate even in the fully connected layer. In contrast, Pong provides dense and frequent reward signals, which may support a cleaner visual-to-abstract transition in the learned hierarchy.

## 7.3 Future Directions

Several improvements could address the limitations identified above and extend the scope of this work.

First, replicating the training using the exact configuration from Cross et al. [3], restoring the original replay buffer size and RMSProp optimizer, would allow direct verification of whether the representational differences are attributable to these modifications, making results more directly comparable with the reference study.

Second, extending the analysis to agents trained with Prioritized Experience Replay (PER) represents a promising direction. PER agents preferentially sample high-error transitions, increasing learning from critical or rare events, a mechanism more consistent with human memory consolidation. For games such as Ms. Pac-Man and Space Invaders, where decisive states are rare, PER could drive the network toward more abstract representations better aligned with human perception.

Third, incorporating electroencephalography (EEG) recordings during task performance would provide a finer-grained, time-resolved measure of neural processing comparable with DQN layer activations using the same RSA framework, bringing the methodology closer to the neural alignment approach of Cross et al. [3] while remaining laboratory-feasible.

Together, these improvements would provide a more principled assessment of the conditions under which human representations of game states align with the hierarchical structure learned by deep reinforcement learning agents.

## REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis - connecting the branches of systems neuroscience," *Frontiers in Systems Neuroscience*, vol. 2, p. 4, 2008.

[3] L. Cross, J. Cockburn, Y. Yue, and J. P. O'Doherty, "Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments," *Neuron*, vol. 109, no. 4, pp. 724–738, 2021.

[4] L. van der Maaten and K. Weinberger, "Stochastic triplet embedding," *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012.

# APPENDIX

## A.1 Behavioral Experiment Links

The sparse experiment is available in two languages:

- Spanish: https://f7pu2mdvua.cognition.run
- English: https://nlelxrroxi.cognition.run

The individual experiment can be accessed at: https://chdkpkrwed.cognition.run

## A.2 Code, Stimuli, and Data

The full source code for the project, including DQN training, stimulus generation, RSA analysis, and experiment scripts, is available at:

https://github.com/Arroch35/TFG-DQNs-Human-Comparison.git

The gameplay clips used as stimuli, the trained DQN model weights, and the anonymised participant response data from both experiments are available at:

http://bit.ly/4uOyKpN

Only trial responses, reaction times, and task metadata are included in the data files; all session identifiers, IP addresses, and timestamps were removed during preprocessing.

## A.3 Triplet Scoring

**Structure Score**

For each triplet $(i, j, k)$, let $d_{\min}$ and $d_{\max}$ denote the smallest and largest of the three pairwise dissimilarities. The structure score is defined as:

$$\sigma = \frac{d_{\max} - d_{\min}}{d_{\max} + \varepsilon} \tag{5}$$

where $\varepsilon$ is a small constant for numerical stability. Triplets with $\sigma$ below the 20th percentile were discarded before difficulty classification.

**Difficulty Score**

For each remaining triplet $(i, j, k)$, where $i$ and $j$ are the similar pair and $k$ is the odd one out, the difficulty score is defined as:

$$\delta = \frac{\bar{d}_{\text{odd}} - d_{\text{similar}}}{\bar{d}_{\text{odd}} + d_{\text{similar}} + \varepsilon} \tag{6}$$

where $d_{\text{similar}}$ is the dissimilarity between clips $i$ and $j$, $\bar{d}_{\text{odd}}$ is the mean dissimilarity from clip $k$ to each of the other two, and $\varepsilon$ is a small constant for numerical stability. Values near 0 indicate hard triplets (the odd clip is only marginally more distant than the similar pair); values near 1 indicate easy triplets (the odd clip is substantially more distant). Triplets were binned into hard (bottom 20%), medium (40th–60th percentile), and easy (top 20%) categories accordingly.

## A.4 Pilot Experiment: Full Results

### A.4.1 Experiment 1: t-STE Embedding Validation

Triplet responses were converted into ordinal constraints and used to train independent t-STE embeddings for each game. Embeddings with 2, 3, 5, and 10 dimensions were evaluated using 6-fold cross-validation across participants. Table 8 summarizes the best-performing dimensionality for each game. Across all games, two dimensions provided the best generalization performance. Pong produced the highest test accuracy, suggesting more consistent similarity judgments across participants. *Ms. Pac-Man* obtained intermediate performance, while *Space Invaders* showed substantially lower generalization accuracy, suggesting that some gameplay clips were more difficult to interpret consistently and motivating later revisions of the stimulus generation process.

| Game | Dim. | Train Acc. | Test Acc. |
|---|---|---|---|
| Pong | 2 | $0.85 \pm 0.03$ | $0.70 \pm 0.11$ |
| Ms. Pac-Man | 2 | $0.84 \pm 0.01$ | $0.61 \pm 0.10$ |
| Space Invaders | 2 | $0.79 \pm 0.01$ | $0.45 \pm 0.08$ |

TABLE 8: Best t-STE cross-validation results for each game.

### A.4.2 Experiment 2: Participant Questionnaire Analysis

The post-experiment questionnaire indicated that participants generally understood the task and felt confident in their decisions. Most participants reported low or moderate difficulty when selecting the odd clip and interpreting the gameplay situations.

Participants showed heterogeneous gaming experience, ranging from no prior gaming experience to frequent gameplay. Nevertheless, most participants reported being familiar with all the Atari games used in the experiment.

Regarding experiment duration, most participants considered the overall length appropriate, although two participants reported that the experiment felt somewhat long. Based on this feedback, reducing the number of triplets per participant was identified as a potential improvement for the deployment version of the experiment.

### A.4.3 Experiment 3: Initial Human-DQN RSA Analysis

Human RDMs were computed from the t-STE embeddings using Euclidean distance. DQN RDMs were computed from activations extracted from convolutional and fully connected layers using correlation distance. RSA was then performed using Spearman correlation between human and DQN RDMs. Initial correlations were inconsistent across games and layers, suggesting potential methodological limitations in the preprocessing and representational analysis pipeline. These observations motivated a detailed revision of the DQN preprocessing procedure and the stimulus generation pipeline described in Appendix A.5.

## A.5 Experiment 4: DQN Representation Pipeline Refinement

### A.5.1 DQN Input Reconstruction

The initial representation extraction procedure used only the final four frames of each gameplay clip to construct the DQN input state. However, the original DQN preprocessing pipeline combines frame skipping and frame stacking operations, as described in Section 4.2.1, requiring a longer temporal sequence to reconstruct the effective network input. Additional frames were included at the beginning of each clip to ensure that the final frame of the reconstructed state corresponded to the same gameplay moment used in the original pilot experiment.

### A.5.2 PCA-based Activation Compression

The initial RSA pipeline computed representational dissimilarity directly from high-dimensional DQN activation vectors, which led to unstable similarity estimates across layers and games. To address this issue, the PCA-based dimensionality reduction procedure described in Section 4.4.2 was incorporated into the pipeline prior to computing the representational dissimilarity matrices.

### A.5.3 Refined RSA Evaluation

Following these pipeline refinements, the RSA analysis was recomputed using the same human RDMs obtained from the pilot experiment. Human and DQN representations were compared independently for each game and network layer using Spearman correlation between RDMs. The resulting RSA values are reported in the section below.

### A.5.4 Pilot RSA Results

Table 9 reports the Spearman correlations between human and DQN RDMs obtained after pipeline refinement, using the pilot experiment human data.

| Game | Layer | RSA ($\rho$) |
|---|---|---|
| Pong | Conv. 1 | -0.084 |
| | Conv. 2 | -0.174 |
| | Conv. 3 | -0.034 |
| | FC | 0.297 |
| Ms. Pac-Man | Conv. 1 | 0.315 |
| | Conv. 2 | 0.126 |
| | Conv. 3 | 0.290 |
| | FC | 0.362 |
| Space Invaders | Conv. 1 | 0.674 |
| | Conv. 2 | 0.464 |
| | Conv. 3 | 0.649 |
| | FC | 0.389 |

TABLE 9: RSA (Spearman correlation) between human and DQN RDMs from the pilot experiment, following pipeline refinement.

In Pong, the highest correlation is observed at the fully connected layer ($\rho = 0.297$), while convolutional layers show near-zero or negative values, suggesting an early signal of human–DQN alignment at the most abstract level of the network. In Ms. Pac-Man, correlations are positive across all layers with relatively similar values, offering no clear layer-specific pattern. In Space Invaders, convolutional layers show generally high correlations ($\rho = 0.464$–$0.674$) while the FC layer drops to $0.389$, though this pattern should be interpreted cautiously given the higher variability in human triplet judgments for this game.

Overall, these results provided an initial signal of alignment between human similarity judgments and DQN representations, while also revealing that the pilot stimulus set and experimental parameters required further refinement before strong conclusions could be drawn, motivating the improvement experiments in Section 5.3.

## A.6 Improvement Experiments

### A.6.1 Optimal Number of Clips: Scaling Results

Table 10 summarizes the mean RSA correlation obtained for different clip subset sizes across the three games.

| Clips | Pong | Ms. Pac-Man | Space Invaders |
|---|---|---|---|
| 10 | $0.84 \pm 0.13$ | $0.69 \pm 0.30$ | $0.70 \pm 0.25$ |
| 12 | $0.88 \pm 0.11$ | $0.75 \pm 0.24$ | $0.74 \pm 0.21$ |
| 15 | $0.92 \pm 0.07$ | $0.80 \pm 0.21$ | $0.80 \pm 0.18$ |
| 20 | $0.95 \pm 0.05$ | $0.86 \pm 0.14$ | $0.85 \pm 0.13$ |
| 25 | $0.96 \pm 0.03$ | $0.89 \pm 0.11$ | $0.87 \pm 0.12$ |
| 30 | $0.97 \pm 0.03$ | $0.92 \pm 0.09$ | $0.90 \pm 0.10$ |
| 50 | $0.98 \pm 0.02$ | $0.95 \pm 0.05$ | $0.92 \pm 0.07$ |

TABLE 10: Mean Spearman correlation between subset RSA structures and the full 909-clip RSA structure for different subset sizes. Results are averaged across 100 random subsets for each condition.

The results show that RSA stability generally increases as the number of clips grows. Using 50 clips produced correlations close to 1.0 for all games, indicating that large subsets are capable of reproducing the layer-wise representational structure of the DQN with high fidelity.

Because the number of possible triplet constraints grows combinatorially with the number of clips, increasing stimulus set size substantially increases the amount of behavioral data required. For this reason, a compromise between representational coverage and experimental feasibility was required.

The results suggest that performance improvements begin to diminish after approximately 15 clips. Based on this trade-off, 15 clips were selected for the behavioral experiment, providing improved representational stability while maintaining a manageable experimental complexity.

### A.6.2 Triplet Reconstruction Analysis: Full Results

Table 11 summarizes the reconstruction performance obtained using different percentages of triplet constraints for the fully connected layer across games, using a two-dimensional t-STE embedding.

Reconstruction quality improved as the number of available triplets increased, with the most notable gains observed between 10% and 40% of the total constraints. After approximately 40% of the triplets, performance gains became

| Triplets | Pong | Ms. Pac-Man | Space Invaders |
|---|---|---|---|
| 10% | $0.56 \pm 0.08$ | $0.77 \pm 0.04$ | $0.87 \pm 0.02$ |
| 20% | $0.63 \pm 0.03$ | $0.83 \pm 0.02$ | $0.89 \pm 0.01$ |
| 40% | $0.67 \pm 0.02$ | $0.86 \pm 0.01$ | $0.89 \pm 0.01$ |
| 60% | $0.68 \pm 0.02$ | $0.86 \pm 0.01$ | $0.89 \pm 0.01$ |
| 80% | $0.68 \pm 0.01$ | $0.87 \pm 0.01$ | $0.90 \pm 0.00$ |
| 100% | $0.69 \pm 0.00$ | $0.87 \pm 0.00$ | $0.90 \pm 0.00$ |

TABLE 11: Mean Spearman correlation between reconstructed t-STE RDMs and original fully connected DQN RDMs for different percentages of sampled triplet constraints, using a two-dimensional embedding. Results are averaged across 100 repetitions.

substantially smaller, suggesting diminishing returns from additional behavioral constraints.

Using all available triplets produced reconstruction correlations of $0.69$, $0.87$, and $0.90$ for Pong, Ms. Pac-Man, and Space Invaders respectively. Ms. Pac-Man and Space Invaders achieved substantially higher reconstruction correlations than Pong, indicating that the two-dimensional embedding captures the FC layer representational geometry of these environments with greater fidelity.

Overall, these results indicate that stable approximations of the original DQN representational geometry can be obtained using only a subset of the total triplet constraints. The experiment therefore provides an estimate of the amount of behavioral data required for the final deployment experiment while maintaining reliable representational reconstructions under the selected two-dimensional embedding.

## A.7 Deployment t-STE Dimensionality Validation

To verify that two dimensions remained sufficient for the deployment data, a dimensionality validation experiment was conducted using the triplet responses collected during the deployment experiment. Embeddings with $2$, $3$, $5$, $8$, $10$, $15$, $20$, and $50$ dimensions were evaluated using Leave-One-Participant-Out (LOPO) cross-validation. For each fold, the embedding was trained on the triplets of all participants except one, and generalization accuracy was evaluated on the held-out participant's triplets. Participants contributing fewer than six triplets for a given game were excluded as test folds but retained in all training sets. The procedure yielded 26, 28, and 25 eligible test folds for Pong, Ms. Pac-Man, and Space Invaders respectively.

Table 12 reports the results. For Ms. Pac-Man and Space Invaders, two dimensions clearly achieved the highest test accuracy. For Pong, three dimensions produced a marginally higher mean test accuracy ($0.691$ vs. $0.670$), but this difference falls well within the standard deviation of both conditions ($\pm 0.16$), and the gap closes further when considering that higher-dimensional embeddings consistently showed increasing training accuracy without a corresponding gain in test accuracy, a pattern indicative of overfitting rather than improved generalization. Given that two dimensions provided the best or statistically equivalent generalization across all three games, and that a uniform embedding dimensionality is required for comparability of the resulting RDMs across games, two dimensions were selected for the final behavioral representational analysis.

| Game | Dim. | Train Acc. | Test Acc. |
|---|---|---|---|
| Pong | 2 | $0.796 \pm 0.009$ | $0.670 \pm 0.162$ |
| | 3 | $0.811 \pm 0.011$ | $0.691 \pm 0.156$ |
| | 5 | $0.840 \pm 0.009$ | $0.654 \pm 0.158$ |
| | 10 | $0.861 \pm 0.009$ | $0.655 \pm 0.153$ |
| | 50 | $0.862 \pm 0.010$ | $0.623 \pm 0.163$ |
| Ms. Pac-Man | 2 | $0.758 \pm 0.009$ | $\mathbf{0.670 \pm 0.171}$ |
| | 3 | $0.765 \pm 0.008$ | $0.601 \pm 0.189$ |
| | 5 | $0.801 \pm 0.009$ | $0.586 \pm 0.167$ |
| | 10 | $0.824 \pm 0.007$ | $0.586 \pm 0.173$ |
| | 50 | $0.830 \pm 0.008$ | $0.569 \pm 0.170$ |
| Space Invaders | 2 | $0.790 \pm 0.013$ | $\mathbf{0.680 \pm 0.171}$ |
| | 3 | $0.811 \pm 0.011$ | $0.637 \pm 0.167$ |
| | 5 | $0.845 \pm 0.010$ | $0.663 \pm 0.171$ |
| | 10 | $0.869 \pm 0.008$ | $0.630 \pm 0.201$ |
| | 50 | $0.878 \pm 0.009$ | $0.633 \pm 0.201$ |

TABLE 12: LOPO cross-validation results for selected dimensionalities in the deployment experiment. Only a representative subset of the evaluated dimensions is shown. Bold indicates the best test accuracy per game; for Pong, the difference between dim=2 and dim=3 does not exceed one standard deviation.

## A.8 DQN Evaluation Performance

Table 13 reports the mean evaluation reward across the five independently trained seeds per game. Rewards follow the standard DQN reward clipping convention, where each positive event is clipped to $+1$ and each negative event to $-1$ per step [1]. The low standard deviation across seeds confirms that all agents reached comparable performance levels, indicating that the representational divergence observed at the FC layer in Experiment 11 is not attributable to differences in learning outcome.

| Game | Mean $\pm$ Std |
|---|---|
| Pong | $15.78 \pm 0.71$ |
| Ms. Pac-Man | $39.28 \pm 2.68$ |
| Space Invaders | $17.19 \pm 1.09$ |

TABLE 13: Mean evaluation reward across five independently trained seeds per game, under standard DQN reward clipping.

## A.9 Continuous DQN Agreement vs. Difficulty

To confirm that the difficulty gradient observed in Experiment 12 is not an artefact of the three-bucket discretisation, cross-seed agreement was recomputed by binning all 455 triplets into ten equal-width intervals along the continuous difficulty score axis ($\delta \in [0, 1]$, bin width $0.1$, minimum three triplets per bin). Table 14 reports the resulting values.

The gradient is monotonically preserved across all three games without exception. The empty intermediate bins for Pong ($\delta \in [0.35, 0.50]$) reflect the bimodal distribution of

| Game | $\delta$ bin | Mean agreement | Std |
|---|---|---|---|
| Pong | 0.05 | 0.469 | 0.017 |
| | 0.15 | 0.683 | 0.040 |
| | 0.25 | 0.770 | 0.053 |
| | 0.55 | 0.979 | 0.014 |
| | 0.65 | 1.000 | 0.000 |
| | 0.95 | 1.000 | 0.000 |
| Ms. Pac-Man | 0.05 | 0.444 | 0.030 |
| | 0.15 | 0.609 | 0.027 |
| | 0.25 | 0.754 | 0.042 |
| | 0.35 | 0.639 | 0.079 |
| | 0.45 | 0.812 | 0.086 |
| | 0.55 | 0.883 | 0.057 |
| | 0.65 | 0.896 | 0.036 |
| Space Invaders | 0.05 | 0.546 | 0.032 |
| | 0.15 | 0.837 | 0.021 |
| | 0.25 | 0.770 | 0.040 |
| | 0.35 | 0.883 | 0.035 |
| | 0.45 | 0.936 | 0.022 |
| | 0.55 | 0.956 | 0.023 |
| | 0.65 | 1.000 | 0.000 |
| | 0.75 | 1.000 | 0.000 |

TABLE 14: DQN cross-seed agreement with the reference seed per difficulty bin and game, using ten equal-width bins along the continuous difficulty score axis. Bins with fewer than three triplets are omitted.

difficulty scores for that game, where triplets cluster at the hard end or jump directly to very easy, with few intermediate cases, consistent with the narrow hard-triplet score range $[0.017, 0.040]$ noted in Experiment 12.

## A.10 Task-Relevant Information in DQN Representations

This experiment examined whether the DQN representations encode task-relevant information in the form of action-value structure. The hypothesis was that the Q-value and state value models would correlate more strongly with deeper layers of the network, where representations are expected to be more closely tied to the agent's policy.

The Q-value model captures similarity between states based on their full action-value profiles as estimated by the DQN. For each state $s_i$, the vector of Q-values over all available actions is extracted from the network:

$$q_i = \big[Q(s_i, a)\big]_{a \in \mathcal{A}} \tag{7}$$

Dissimilarity between two states is computed using correlation distance between their Q-value vectors:

$$d_{ij}^{Q} = 1 - corr(q_i, q_j) \tag{8}$$

The state value model captures similarity between states based on their scalar expected return as estimated by the policy. The value of a state is defined as the maximum Q-value over all actions:

$$V(s_i) = \max_a Q(s_i, a) \tag{9}$$

Dissimilarity between two states is computed as the absolute difference between their values:

$$d_{ij}^{V} = |V(s_i) - V(s_j)| \tag{10}$$

Table 15 reports the RSA correlations between both models and each DQN layer across games.

| Game | Layer | Q-value ($\rho$) | State Value ($\rho$) |
|---|---|---|---|
| Pong | Conv. 1 | 0.111 | 0.004 |
| | Conv. 2 | 0.058 | 0.206 |
| | Conv. 3 | -0.054 | 0.249 |
| | FC | -0.151 | -0.023 |
| Ms. Pac-Man | Conv. 1 | 0.168 | 0.099 |
| | Conv. 2 | 0.023 | -0.014 |
| | Conv. 3 | 0.208 | 0.168 |
| | FC | 0.087 | 0.074 |
| Space Invaders | Conv. 1 | 0.184 | 0.179 |
| | Conv. 2 | 0.155 | 0.105 |
| | Conv. 3 | 0.166 | 0.142 |
| | FC | 0.162 | 0.153 |

TABLE 15: RSA correlations (Spearman $\rho$) between the Q-value and state value RDMs and each DQN layer, across the three games.

Neither model produced meaningful correlations with any DQN layer in any of the three games. In Ms. Pac-Man, the Q-value and state value RDMs showed near-perfect mutual correlation ($\rho = 0.968$), indicating that both models capture essentially the same information for this game, yet neither aligned with the network representations. No layer showed a consistently higher alignment across games, and the expected gradient toward deeper layers was not observed. These models are therefore not considered further in the analysis.

## A.11 Full RSA Matrices

Figure 1 shows the full RSA matrices for all three games, reporting the Spearman correlation between every pair of RDMs computed in the analysis: DQN layers, pixel model, pixel PCA model, Q-value model, state value model, HCF model (Pong only), and the human sparse RDM.

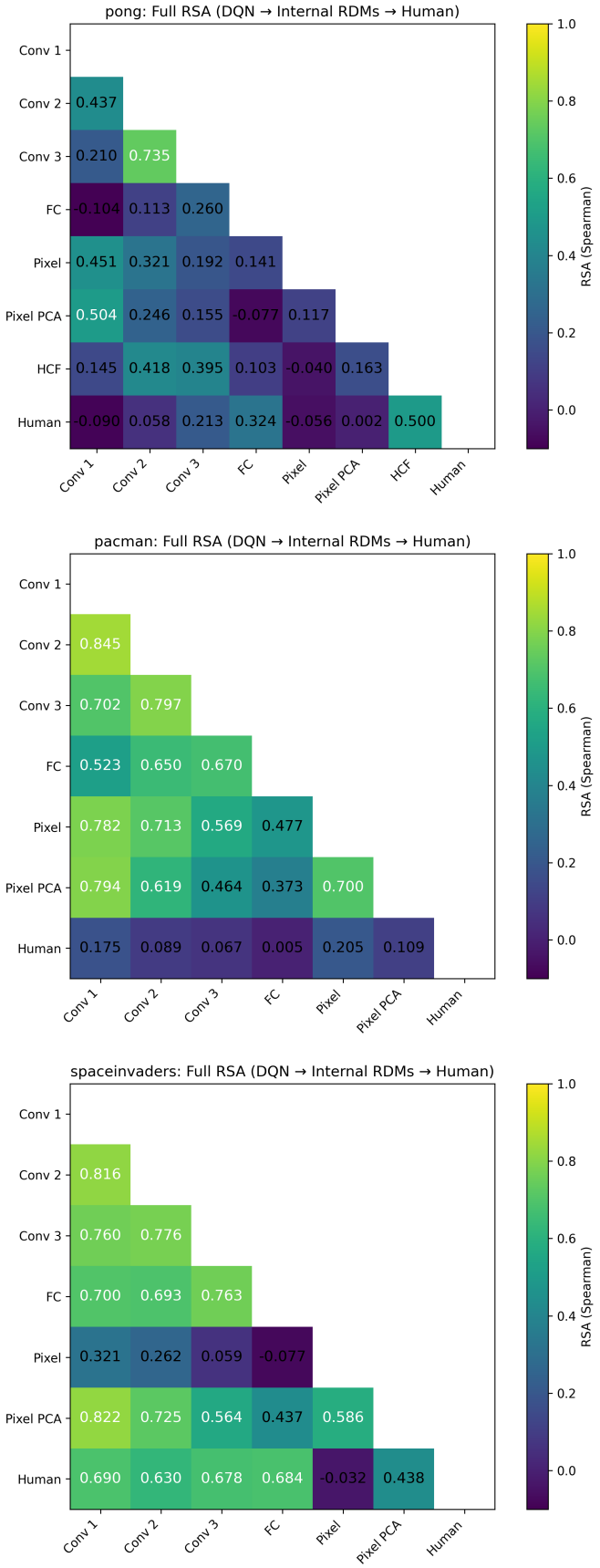



Fig. 1: Full RSA matrices for Pong (left), Ms. Pac-Man (center), and Space Invaders (right). Each cell reports the Spearman correlation between a pair of RDMs: DQN layers (Conv. 1–Conv. 3, FC), pixel model, pixel PCA model, Q-value model, state value model, HCF model (Pong only), and the human sparse RDM. Warmer colors indicate higher representational similarity.