



**Εθνικό Μετσόβιο Πολυτεχνείο  
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών  
Υπολογιστών**

Εξάμηνο : 9ο

Μάθημα : Νευρο-ασαφής Έλεγχος και Εφαρμογές

Διδάσκων : I. Κορδώνης

Ακαδημαϊκό έτος : 2025-26

**Εξαμηνιαία εργασία**

**Google Dreamer V1**

Όναματεπώνυμο: Μωραΐτης Νικόλαος

Αριθμός Μητρώου: 031 21172

## Περιεχόμενα

---

<b>1 Περίληψη</b>	<b>2</b>
<b>2 Εισαγωγή</b>	<b>3</b>
2.1 Reinforcement Learning . . . . .	3
2.2 Google Dreamer . . . . .	3
2.3 Model-Based vs Model-Free RL . . . . .	4
<b>3 Components</b>	<b>5</b>
3.1 Variational Autoencoder (VAE) . . . . .	5
3.1.1 Encoder: $p(s_t   s_{t-1}, a_{t-1}, o_t)$ . . . . .	5
3.1.2 Decoder: $p(o_t   s_t)$ . . . . .	6
3.1.3 Επιλογή Ασπρόμαυρης Επεξεργασίας . . . . .	9
3.2 Recurrent State Space Model (RSSM) . . . . .	9
3.3 Reward Model . . . . .	10
3.4 Actor . . . . .	10
3.4.1 Actor Structure . . . . .	10
3.4.2 Reparameterization Trick . . . . .	11
3.5 Critic . . . . .	11
3.6 Value Estimation ( $V_\lambda$ ) . . . . .	12
<b>4 Αλγόριθμος</b>	<b>13</b>
4.1 Αρχικοποίηση . . . . .	13
4.2 Φάση 1: Συλλογή δεδομένων . . . . .	14
4.3 Φάση 2: Representation Learning . . . . .	14
4.4 Φάση 3: Χώρος Λανθάνουσας Φαντασίας - Εκπαίδευση Συμπεριφοράς	15
<b>5 Αποτελέσματα</b>	<b>17</b>
5.1 Standard Ρύθμιση . . . . .	17
5.2 Σύνοψη Εκπαίδευσης . . . . .	17
5.3 Σύγκριση Επιδόσεων (Performance Comparison) . . . . .	19
5.4 Οπτική Παρατήρηση σε διάφορα στάδια εκπαίδευσης . . . . .	20
<b>6 Συμπεράσματα</b>	<b>25</b>

---

## 1 Περίληψη

---

Στόχος αυτής της εργασίας ήταν η κατανόηση και υλοποίηση του Google Dream V1. Τί είναι το google dreamer; Το google dreamer είναι ένας αλγόριθμος ενισχυτικής μάθησης που χρησιμοποιεί τον λανθάνον χώρο για εκπαίδευση. Αποτελείται από έναν Variational Auto Encoder (VAE) τον οποίο έχουμε σπάσει σε Encoder και Decoder και ένα RSSM (Recurrent State Space Model). Στόχος του είναι να πραγματοποιήσει έλεγχο σε ένα σύστημα όπως τετράποδο ρομπότ ανάποδο εκκρεμές και άλλα συστήματα. Εμείς θα εκπαιδεύσουμε το Dreamer στο παιχνίδι του gymnasium CarRacing-V3 όπου στόχος είναι το αυτοκίνητο να παραμείνει εντός ενός δρόμου για όσο το δυνατό περισσότερο. Εκπαιδεύουμε για όσο το δυνατόν περισσότερα βήματα, χωρίς gpus και τοπικά και στο colab. Για λόγους οικονομίας στην εκπαίδευση βάλαμε ασπρόμαυρες εικόνες αντί για έγχρωμες.

## 2 Εισαγωγή

---

### 2.1 Reinforcement Learning

Η ενισχυτική μάθηση (Reinforcement Learning, RL) αποτελεί μια περιοχή της τεχνητής νοημοσύνης που εστιάζει στην εκμάθηση συμπεριφορών μέσα από την αλληλεπίδραση ενός πράκτορα (agent) με ένα περιβάλλον. Ο πράκτορας επιλέγει ενέργειες έχοντας ως στόχο τη μεγιστοποίηση της συνολικής επιβράβευσης, χρησιμοποιώντας εμπειρία που συλλέγεται κατά τη διάρκεια της εκπαίδευσης. Παρά την αποτελεσματικότητά της σε απλά περιβάλλοντα, η ενισχυτική μάθηση συχνά δυσκολεύεται σε περιβάλλοντα υψηλής διάστασης, όπου οι παρατηρήσεις (π.χ. εικόνες) περιέχουν μεγάλο όγκο περιττής πληροφορίας.

Η ενισχυτική μάθηση μοντελοποιεί τη διαδικασία λήψης αποφάσεων ως ένα Markov Decision Process (MDP), ορίζοντας: State ( $s$ ): Η τρέχουσα κατάσταση του περιβάλλοντος η οποία, στο δικό μας σύστημα περιλαμβάνει ένα ντετερμινιστικό κομμάτι  $h$  και ένα στοχαστικό κομμάτι  $z$ . Το ντετερμινιστικό κομμάτι εκφράζει την προβλέψιμη δυναμική του συστήματος. Το στοχαστικό κομμάτι αντιπροσωπεύει αβεβαιότητες και επιτρέπει την πρόβλεψη πιθανών μελλοντικών σεναρίων. Action ( $a$ ): Η ενέργεια που εκτελεί ο πράκτορας, π.χ. επιτάχυνση ή στροφή του αυτοκινήτου. Reward ( $r$ ): Η επιβράβευση που λαμβάνει ο πράκτορας για κάθε ενέργεια. Policy ( $\pi$ ): Στρατηγική επιλογής ενεργειών βάσει της κατάστασης. Value function ( $V$ ): Εκτιμά το συνολικό αναμενόμενο reward από μία κατάσταση. Στόχος του πράκτορα είναι η μεγιστοποίηση του συνολικού αναμενόμενου reward μέσω της εκπαίδευσης της policy του.

### 2.2 Google Dreamer

Ο αλγόριθμος του Google Dreamer V1 που προτάθηκε από τους Hafner et al. το 2020 αποτελεί μία σύγχρονη προσέγγιση ενισχυτικής μάθησης που προσπαθεί να αντιμετωπίσει αυτές τις προκλήσεις. Το Dreamer παίρνει έμπνευση από τα μοντέλα λανθάνοντος χώρου και τα συνδυάζει με actor-critic πολιτικές, επιτρέποντας στον agent να ονειρεύεται μελλοντικές καταστάσεις και να εκπαιδεύεται σε έναν λανθάνοντα χώρο αντί σε ακριβείς πραγματικές παρατηρήσεις. Αυτό επιτρέπει την ταχύτερη και λιγότερο κοστοβόρα εκπαίδευση αποφεύγοντας την επανεκκίνηση του χώρου προσομοίωσης σε κάθε βήμα. Το οποίο είναι ιδιαίτερα χρήσιμο σε περιβάλλοντα με υψηλές διαστάσεις όπως τα παιχνίδια με εικόνες από τις οποίες πρέπει να εξάγουμε

μόνο την χρήσιμη πληροφορία. Στην παρούσα εργασία υλοποιούμε από την αρχή και αναλύουμε το google dreamer V1 στο περιβάλλον CarRacing-V3 του gymnasium. Στόχος είναι η εκπαίδευση ενός πράκτορα που ελέγχει το αυτοκίνητο ώστε να παραμένει εντός του δρόμου για όσο το δυνατόν περισσότερο χρόνο.

### 2.3 Model-Based vs Model-Free RL

Σε αντίθεση με αλγορίθμους Model-Free (όπως PPO, DQN) που μαθαίνουν απευθείας από την αλληλεπίδραση, το Dreamer (Model-Based) προσφέρει:

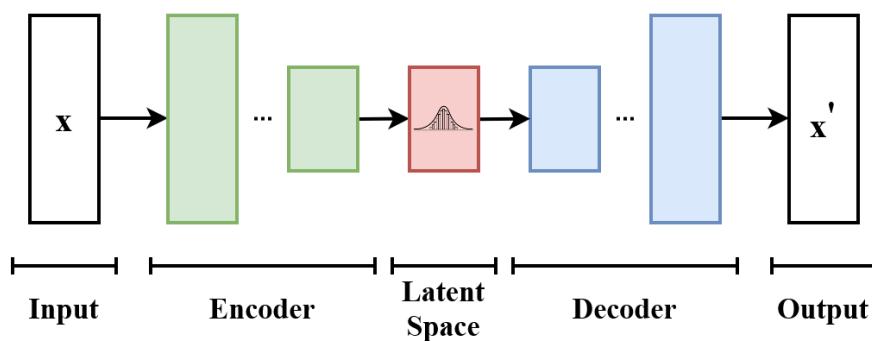
- **Sample Efficiency:** Χρειάζεται λιγότερα βήματα στο περιβάλλον, καθώς εκπαιδεύεται χιλιάδες φορές πάνω στα ίδια δεδομένα εντός της "φαντασίας" του.
- **Long-horizon Prediction:** Μέσω του RSSM, μπορεί να προβλέψει τις συνέπειες ενεργειών πολύ πιο μακριά στο μέλλον από ό,τι ένας απλός Q-learning αλγόριθμος.

## 3 Components

Ας δούμε όλα τα components ένα ένα:

### 3.1 Variational Autoencoder (VAE)

Ο VAE αποτελεί ένα βασικό εργαλείο για την κωδικοποίηση εικόνων σε λανθάνοντα χώρο. Περιλαμβάνει έναν encoder και έναν decoder



**Σχήμα 1:** Βασική δομή ενός VAE [5]

#### 3.1.1 Encoder: $p(s_t | s_{t-1}, a_{t-1}, o_t)$

Ο ρόλος του είναι να παίρνει μία εικόνα και να την μετατρέπει σε μοντέλο. Το να επεξεργαζόμαστε συνεχώς μία εικόνα δεν έχει νόημα καθώς εμπεριέχει πολλή περιττή πληροφορία και βαραίνει την εκπαίδευση. Αντίθετα αυτό που χρειαζόμαστε είναι να εξάγουμε την δυναμική του μοντέλου. Αντί να την εξάγουμε μέσω φυσικής την προσεγγίζουμε με την χρήση του encoder [4].

- Είσοδος: 64x64 (ασπρόμαυρη εικόνα)
- Έξοδος: Ένα διάνυσμα με μέγεθος 1024 bits

```

1 self.net = nn.Sequential(
2     # (1, 64, 64) -> (32, 31, 31)
3     nn.Conv2d(in_channels, 32, kernel_size=4, stride=2),
4     nn.ReLU(),
5
6     # (32, 31, 31) -> (64, 14, 14)
7     nn.Conv2d(32, 64, kernel_size=4, stride=2),

```

```

8     nn.ReLU(),
9
10    # (64, 14, 14) -> (128, 6, 6)
11    nn.Conv2d(64, 128, kernel_size=4, stride=2),
12    nn.ReLU(),
13
14    # (128, 6, 6) -> (256, 2, 2)
15    nn.Conv2d(128, 256, kernel_size=4, stride=2),
16    nn.ReLU(),
17
18    # (256, 2, 2) -> (1024, )
19    nn.Flatten()
20
21
22 # Output dimension
23 self.embed_dim = 256 * 2 * 2 # = 1024

```

### 3.1.2 Decoder: $p(o_t | s_t)$

Ο ρόλος του είναι να λάβει ένα state  $[h, z]$  και να την μετατρέπει σε ασπρόμαυρη εικόνα  $\hat{o}_t$ . Μας είναι χρήσιμος καθώς μας επιτρέπει να απεικονίσουμε την φαντασία και να έχουμε μία καλύτερη εποπτεία με μια “γενικευμένη εικόνα” του δρόμου.

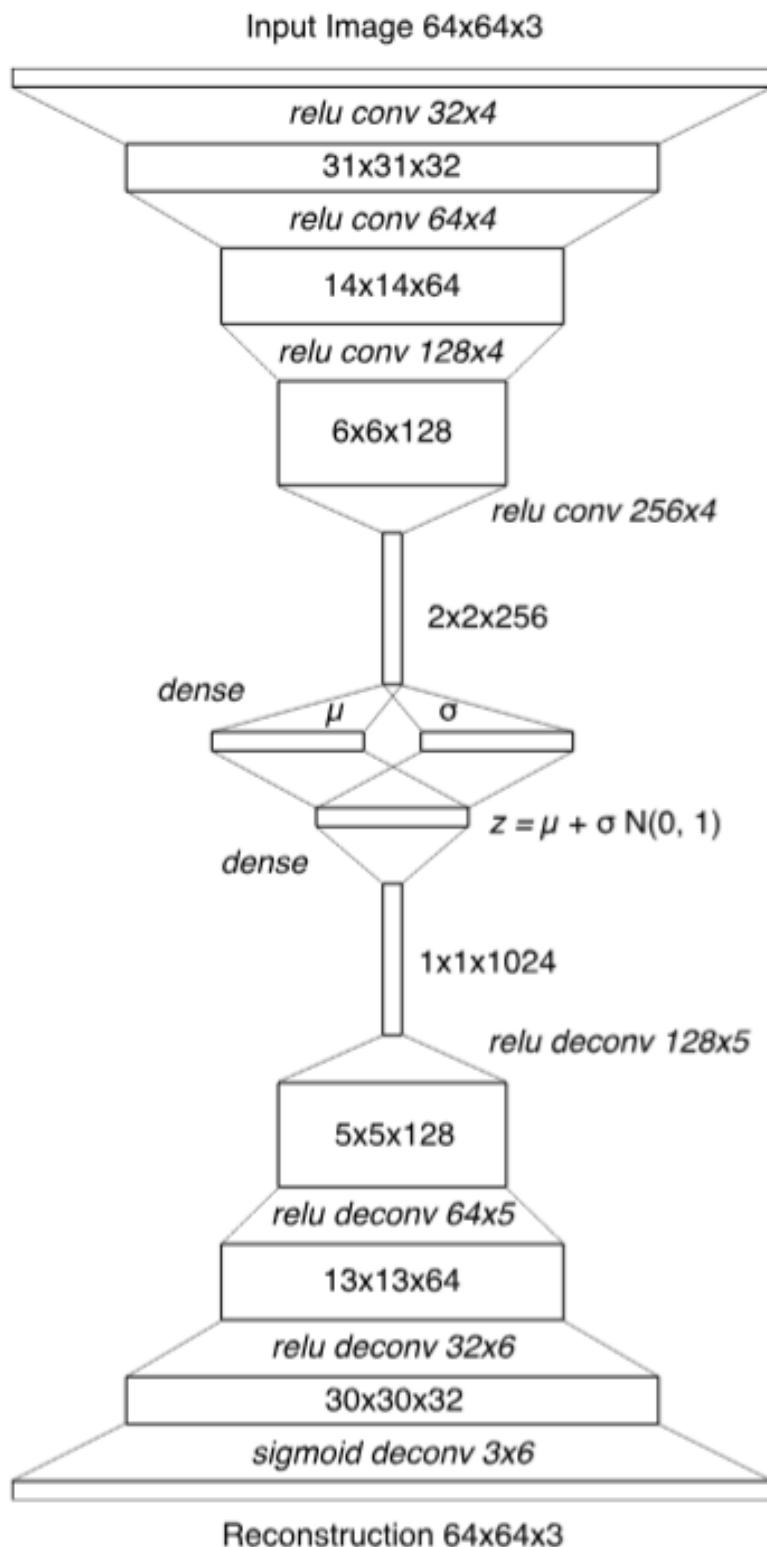
- Είσοδος: 230 αριθμοί
- Έξοδος: 64x64 (ασπρόμαυρη εικόνα)

```

1 def __init__(self, state_dim=230, out_channels=1):
2     """
3
4     Args:
5         state_dim: Dimension of model state (h, z). Default: 200 + 30 = 230
6         out_channels: Number of output image channels (1 for grayscale)
7     """
8
9     super().__init__()
10
11    # Project state to spatial format
12    self.fc = nn.Linear(state_dim, 1024)
13
14    # Transposed convolutions to upsample
15    self.net = nn.Sequential(

```

```
14     # (1024,) -> (1024, 1, 1)
15     nn.Unflatten(1, (1024, 1, 1)),
16
17     # (1024, 1, 1) -> (128, 5, 5)
18     nn.ConvTranspose2d(1024, 128, kernel_size=5, stride=2),
19     nn.ReLU(),
20
21     # (128, 5, 5) -> (64, 13, 13)
22     nn.ConvTranspose2d(128, 64, kernel_size=5, stride=2),
23     nn.ReLU(),
24
25     # (64, 13, 13) -> (32, 30, 30)
26     nn.ConvTranspose2d(64, 32, kernel_size=6, stride=2),
27     nn.ReLU(),
28
29     # (32, 30, 30) -> (1, 64, 64)
30     nn.ConvTranspose2d(32, out_channels, kernel_size=6, stride=2),
31     nn.Sigmoid() # Output in [0, 1] range
32 )
```



**Σχήμα 2:** Η αναλυτική δομή του VAE που χρησιμοποιείται στον Dreamer [3]

### 3.1.3 Επιλογή Ασπρόμαυρης Επεξεργασίας

Η επιλογή μετατροπής των εικόνων σε ασπρόμαυρες (grayscale) αντί για έγχρωμες (RGB) έγινε για τους εξής λόγους:

- **Μείωση υπολογιστικού κόστους:** Μία ασπρόμαυρη εικόνα  $64 \times 64$  έχει 4,096 τιμές, ενώ μία έγχρωμη έχει 12,288 ( $3 \times$  περισσότερες). Αυτό μειώνει σημαντικά τον χρόνο εκπαίδευσης, ιδιαίτερα χωρίς GPU.
- **Επαρκής πληροφορία:** Στο περιβάλλον CarRacing-V3, η κρίσιμη πληροφορία (θέση δρόμου, όρια πίστας, θέση αυτοκινήτου) διατηρείται πλήρως στην ασπρόμαυρη αναπαράσταση. Το χρώμα δεν προσφέρει επιπλέον σημασιολογική πληροφορία για την εργασία.
- **Μικρότερο μοντέλο:** Ο encoder χρειάζεται λιγότερες παραμέτρους (1 κανάλι εισόδου αντί για 3), μειώνοντας την πιθανότητα overfitting με περιορισμένα δεδομένα εκπαίδευσης.
- **Συμβατότητα με World Models:** Η αρχιτεκτονική VAE που χρησιμοποιούμε βασίζεται στην εργασία των Ha & Schmidhuber [3], η οποία επίσης χρησιμοποιεί ασπρόμαυρες εικόνες.

## 3.2 Recurrent State Space Model (RSSM)

Αποτελείται από 2 μέρη. Μία ντετερμινιστική κατάσταση  $h$  και μία στοχαστική κατάσταση  $z$  που αντιπροσωπεύει την αβεβαιότητα της παρούσας κατάστασης. Ο συνδυασμός μας επιτρέπει να προβλέπουμε πιθανά σενάρια όχι μόνο τα σίγουρα.

- Η λειτουργία **Transition** του RSSM πραγματοποιεί την μετάβαση  $q(s_t | s_{t-1}, a_{t-1})$ . Πηγαίνει στην λανθάνουσα κατάσταση  $h_t = \text{GRU}(h_{t-1}, \text{concat}(z_{t-1}, a_{t-1}))$  που είναι η πρόβλεψη για το μέλλον. Στη συνέχεια υπολογίζει την πιθανότητα της επόμενης κατάστασης με βάση την παρούσα δηλαδή την prior πιθανότητα  $p(z_t | h_t)$ . στον κώδικα το Transition υλοποιείται από την συνάρτηση `imagine`.
- Η λειτουργία **Observe** πραγματοποιεί τον υπολογισμό  $p(s_t | s_{t-1}, a_{t-1}, o_t)$ . Δηλαδή, μετά το όνειρο, εκτιμάται η posterior πιθανότητα  $q(z_t | h_t, e_t)$  όπου  $e_t$  το σφάλμα μεταξύ  $z_t$  και  $h_t$ . Μέσω αυτής ενημερώνεται το μοντέλο.

Στον πυρήνα του RSSM βρίσκεται ένα multi-layer Gated Recurrent Unit (GRU) RNN.

### 3.3 Reward Model

Ο ρόλος του είναι να λάβει την παρούσα κατάσταση  $[h, z]$  και να προβλέψει την επιβράβευση της κατάστασης δηλαδή υλοποιεί την  $q(r_t \mid s_t)$ . Αυτό πραγματοποιείται μέσω ενός MLP.

- **Είσοδος:**  $[h, z]$  ( $\text{len}([h, z]) = 230$  αριθμοί)
- **Έξοδος:** Ένας αριθμός

```

1 def __init__(self, state_dim, hidden_dim=300):
2     super().__init__()
3     self.net = nn.Sequential(
4         nn.Linear(state_dim, hidden_dim),
5         nn.ELU(),
6         nn.Linear(hidden_dim, hidden_dim),
7         nn.ELU(),
8         nn.Linear(hidden_dim, 1)
9     )

```

### 3.4 Actor

Ο ρόλος του είναι να αποφασίσει τι δράση να κάνει σε κάθε περίπτωση. Με βάση την αναμενόμενη τιμή του κέρδους αποφασίζει ποια είναι η καλύτερη επόμενη κίνηση. Ωστόσο προσθέτουμε και μία τυχαιότητα στις κινήσεις με στόχο την εξερεύνηση ενός μεγαλύτερου χώρου κινήσεων. Για να μετατρέψουμε την κίνηση σε actuation χρησιμοποιούμε την `tanh` και κάνουμε squash τις τιμές της επιτάχυνσης (γκάζι, φρένο) και στροφής στο χωρίο  $[-1, 1]$ .

#### 3.4.1 Actor Structure

```

1 def __init__(self, state_dim, action_dim, hidden_dim=300, min_std=0.1, init_std
2 =5.0):
3     super().__init__()
4     self.min_std = min_std
5     self.init_std = init_std
6
7     self.net = nn.Sequential(
8         nn.Linear(state_dim, hidden_dim),
9

```

```

8     nn.ELU(),
9     nn.Linear(hidden_dim, hidden_dim),
10    nn.ELU(),
11 )
12
13 self.mean_head = nn.Linear(hidden_dim, action_dim)
14 self.std_head = nn.Linear(hidden_dim, action_dim)

```

### 3.4.2 Reparameterization Trick

Μία κρίσιμη λεπτομέρεια για την εκπαίδευση του VAE και του Actor είναι η δυνατότητα υπολογισμού παραγώγων μέσω στοχαστικών κόμβων. Κανονικά, η δειγματοληψία  $z \sim \mathcal{N}(\mu, \sigma)$  δεν είναι παραγωγίσιμη πράξη. Ο Dreamer χρησιμοποιεί το **Reparameterization Trick**:

$$z = \mu + \sigma \cdot \epsilon, \quad \text{όπου } \epsilon \sim \mathcal{N}(0, I) \quad (1)$$

Αυτό επιτρέπει στα gradients να περάσουν ανεμπόδιστα προς τα  $\mu$  και  $\sigma$ , επιτρέποντας την εκπαίδευση του Encoder και του Actor μέσω backpropagation through time (BPTT).

## 3.5 Critic

Ο κριτής έχει ως στόχο την παραγωγή μίας εκτίμησης για το συνολικό κέρδος μέχρι τον ορίζοντα των 15 βημάτων.

```

1 def __init__(self, state_dim, hidden_dim=300):
2     super().__init__()
3     self.net = nn.Sequential(
4         nn.Linear(state_dim, hidden_dim),
5         nn.ELU(),
6         nn.Linear(hidden_dim, hidden_dim),
7         nn.ELU(),
8         nn.Linear(hidden_dim, 1)
9     )

```

### 3.6 Value Estimation ( $V_\lambda$ )

Για να εξισορροπηθεί η μεροληψία (bias) και η διακύμανση (variance), ο Dreamer χρησιμοποιεί τον εκτιμητή  $V_\lambda$ , ο οποίος είναι ένας εκθετικά σταθμισμένος μέσος όρος των εκτιμήσεων για διαφορετικούς ορίζοντες  $k$ .

$$V_R(s_\tau) \doteq \mathbb{E}_{q_\theta, q_\phi} \left( \sum_{n=\tau}^{t+H} r_n \right), \quad (4)$$

$$V_N^k(s_\tau) \doteq \mathbb{E}_{q_\theta, q_\phi} \left( \sum_{n=\tau}^{h-1} \gamma^{n-\tau} r_n + \gamma^{h-\tau} v_\psi(s_h) \right) \quad \text{with} \quad h = \min(\tau + k, t + H), \quad (5)$$

$$V_\lambda(s_\tau) \doteq (1 - \lambda) \sum_{n=1}^{H-1} \lambda^{n-1} V_N^n(s_\tau) + \lambda^{H-1} V_N^H(s_\tau), \quad (6)$$

## 4 Αλγόριθμος

Η διαδικασία που ακολουθεί το Dreamer κατά την διάρκεια της εκπαίδευσης είναι η εξής:

**Algorithm 1:** Dreamer

<b>Model components</b>	
Representation	$p_\theta(s_t   s_{t-1}, a_{t-1}, o_t)$
Transition	$q_\theta(s_t   s_{t-1}, a_{t-1})$
Reward	$q_\theta(r_t   s_t)$
Action	$q_\phi(a_t   s_t)$
Value	$v_\psi(s_t)$
<b>Hyper parameters</b>	
Seed episodes	$S$
Collect interval	$C$
Batch size	$B$
Sequence length	$L$
Imagination horizon	$H$
Learning rate	$\alpha$

```

Initialize dataset  $\mathcal{D}$  with  $S$  random seed episodes.
Initialize neural network parameters  $\theta, \phi, \psi$  randomly.
while not converged do
    for update step  $c = 1..C$  do
        // Dynamics learning
        Draw  $B$  data sequences  $\{(a_t, o_t, r_t)\}_{t=k}^{k+L} \sim \mathcal{D}$ .
        Compute model states  $s_t \sim p_\theta(s_t | s_{t-1}, a_{t-1}, o_t)$ .
        Update  $\theta$  using representation learning.
        // Behavior learning
        Imagine trajectories  $\{(s_\tau, a_\tau)\}_{\tau=t}^{t+H}$  from each  $s_t$ .
        Predict rewards  $E(q_\theta(r_\tau | s_\tau))$  and values  $v_\psi(s_\tau)$ .
        Compute value estimates  $V_\lambda(s_\tau)$  via Equation 6.
        Update  $\phi \leftarrow \phi + \alpha \nabla_\phi \sum_{\tau=t}^{t+H} V_\lambda(s_\tau)$ .
        Update  $\psi \leftarrow \psi - \alpha \nabla_\psi \sum_{\tau=t}^{t+H} \frac{1}{2} \|v_\psi(s_\tau) - V_\lambda(s_\tau)\|^2$ .
        // Environment interaction
         $o_1 \leftarrow \text{env.reset}()$ 
        for time step  $t = 1..T$  do
            Compute  $s_t \sim p_\theta(s_t | s_{t-1}, a_{t-1}, o_t)$  from history.
            Compute  $a_t \sim q_\phi(a_t | s_t)$  with the action model.
            Add exploration noise to action.
             $r_t, o_{t+1} \leftarrow \text{env.step}(a_t)$ .
            Add experience to dataset  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(o_t, a_t, r_t)\}_{t=1}^T$ .
    
```

Ποιο αναλυτικά και διαισθητικά, η παραπάνω διαδικασία περιγράφεται παρακάτω:

### 4.1 Αρχικοποίηση

Αρχικοποιούμε το βέλτιστο μέσο reward σε  $-\infty$ , τις παραμέτρους του μοντέλου ( $\theta$ ) και του κριτή ( $\psi$ ) τυχαία και παράγουμε  $S = 5$  επεισόδια. Επιπλέον θέτουμε τις υπερπαραμέτρους ως εξής:

- Until done = 500.000 βήματα
- $C = 100, B = 50, L = 50, H = 15, T = 50$
- **Learning rates:**
  - World model (encoder, RSSM, decoder, reward):  $6 \times 10^{-4}$
  - Value model (critic):  $8 \times 10^{-5}$
  - Action model (actor):  $8 \times 10^{-5}$

## 4.2 Φάση 1: Συλλογή δεδομένων

1. **Προεπεξεργασία:** Μετατροπή του frame  $o_t$  σε ασπρόμαυρο, κανονικοποίηση των τιμών.
2. Ο encoder μετατρέπει το  $o_t$  σε εκτίμηση της κατάστασης του συστήματος  $[h, z]$ .
3. Το RSSM παρατηρεί την τωρινή κατάσταση και εκτιμά τις prior και posterior της προηγούμενης κατάστασης.
4. Ο Actor επιλέγει μία κίνηση (action) με βάση την τωρινή κατάσταση.
5. Προσθέτουμε θόρυβο στο action για να επιτρέψουμε την εξερεύνηση νέων πολιτικών.
6. Πραγματοποιούμε την κίνηση στον πραγματικό κόσμο και παίρνουμε την επιβράβευση μας καθώς και το επόμενο frame  $o_{t+1}$ .
7. Αποθηκεύουμε την νέα κίνηση στον Replay Buffer.
8. Επαναλαμβάνουμε τα βήματα 1-8 μέχρι τις  $C$  επαναλήψεις.

## 4.3 Φάση 2: Representation Learning

1. Παίρνουμε ένα σύνολο κινήσεων από τον replay buffer (batch).
2. Αρχικοποιούμε τις  $T$  καταστάσεις του batch.
3. Πραγματοποιούμε ένα βήμα εκπαίδευσης  $T$  φορές:
  - (α') Παίρνουμε το frame και το κωδικοποιούμε ως  $e_t$ .
  - (β') Το RSSM γνωρίζοντας την προηγούμενη κατάσταση, το προηγούμενο action και το τωρινό encoding παράγει την prior, posterior και το  $[h, z] = s_t$ .
  - (γ') Πραγματοποιούμε decode του παρόντος state  $s_t$  και βρίσκουμε το σφάλμα του Decoder με βάση την πραγματική εικόνα  $o_t$ . Το σφάλμα υπολογίζεται με MSE.
  - (δ') Προβλέπουμε το reward του  $s_t$  που προβλέψαμε  $\hat{r}_t$  και το συγκρίνουμε με το αποθηκευμένο στον replay buffer reward  $r_t$ . Βρίσκουμε MSE.
  - (ε') Υπολογίζουμε  $KL(\text{prior} \parallel \text{posterior})$  όπου  $KL = \text{Kullback-Leibler divergence}$ .
4. Με βάση τα παραπάνω υπολογίζουμε το συνολικό σφάλμα ως εξής:

$$\text{total\_loss} = \text{reconstruction\_loss} + \text{reward\_loss} + \beta \times \text{kl\_loss}$$

5. Πραγματοποιούμε back propagation και στη συνέχεια ενημερώνουμε τις παραμέτρους

του μοντέλου  $\theta$  (world\_params) χρησιμοποιώντας Adam optimizer. Έτσι τελειώνει το μέρος της εκπαίδευσης του μοντέλου.

#### 4.4 Φάση 3: Χώρος Λανθάνουσας Φαντασίας - Εκπαίδευση Συμπεριφοράς

Στη συνέχεια μπαίνουμε στον λανθάνοντα χώρο. Παίρνουμε μία κίνηση από το ιστορικό (η οποία είχε παραχθεί από τον Actor) και το RSSM προσπαθεί να «ονειρευτεί» την επόμενη κατάσταση όπως στο 11β. Επαναλαμβάνουμε τα παρακάτω  $T$  φορές:

1. Όνειρο και εύρεση  $[h_\tau, z_\tau]$  ( $\tau$  συμβολίζει τον λανθάνοντα χρόνο).
2. Το Reward model προβλέπει το reward της επόμενης κατάστασης.
3. Αποθήκευση κατάστασης και reward  $r_\tau$ .
4. Ο Critic προβλέπει το reward μέχρι το τέλος του ορίζοντα ( $H$  επαναλήψεις) για κάθε ένα από τα  $T$  states που φανταστήκαμε (έχουν παραχθεί από τον Actor).
5. Υπολογίζουμε το  $V_\lambda$  χρησιμοποιώντας τα  $r_\tau$  και  $v_\psi(s_\tau)$ .
6. Υπολογίζουμε τα  $v_\psi(s_\tau)$  ξανά με gradients.
7. Πραγματοποιούμε back propagation στο actor\_loss =  $-V_\lambda.\text{mean}$ .
8. Πραγματοποιούμε Adam optimization στον actor ενημερώνοντας την πολιτική του.
9. Κάνουμε detach στα imagined states, δηλαδή για  $T \times H$  καταστάσεις παγώνοντας τις επιλογές που έγιναν αφού δεν θέλουμε να τις αλλάξουμε.
10. Συγκρίνουμε με τις τιμές  $v_\psi$  με τα  $V_\lambda$  που βρήκαμε στο 16 και βρίσκουμε το MSE.
11. Πραγματοποιούμε backpropagation στο σφάλμα MSE του κριτή.
12. Ενημερώνουμε τον τρόπο εκτίμησης επιβράβευσης του Critic χρησιμοποιώντας Adam optimizer.
13. Επαναλαμβάνουμε τις 3 παραπάνω φάσεις για κάθε βήμα εκπαίδευσης με σύνολο 23.000 βήματα (Η google έκανε 5.000.000).

**Σημασία του KL Divergence:** Ο όρος  $L_{KL}$  στη συνάρτηση κόστους εξυπηρετεί δύο σκοπούς:

1. Λειτουργεί ως regularization, αποτρέποντας τον λανθάνοντα χώρο από το να γίνει ασυνεχής (κάτι που θα δυσκόλευε το "όνειρο").
2. Εξασφαλίζει ότι η posterior κατανομή  $q(z_t | \cdot)$  παραμένει κοντά στην prior  $p(z_t | \cdot)$ , επιτρέποντας στο μοντέλο να γενικεύει σε νέες καταστάσεις.

**Ορισμός KL Divergence:** Η απόκλιση Kullback-Leibler ( $D_{KL}$ ) είναι ένα μέτρο της διαφοράς μεταξύ δύο κατανομών πιθανότητας  $P$  και  $Q$ . Για συνεχείς τυχαίες μεταβλητές, ορίζεται ως το ολοκλήρωμα:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \left( \frac{p(x)}{q(x)} \right) dx \quad (2)$$

Ουσιαστικά, εκφράζει την "πληροφοριακή απόσταση" που χάνεται αν χρησιμοποιήσουμε την κατανομή  $Q$  για να προσεγγίσουμε την πραγματική κατανομή  $P$ .

**KL Divergence για Κανονικές Κατανομές:** Στον αλγόριθμο Dreamer, όπου οι κατανομές posterior  $q(z|x)$  και prior  $p(z)$  είναι πολυμεταβλητές Γκαουσιανές (Multivariate Gaussians), η KL Divergence υπολογίζεται αναλυτικά ως:

$$D_{KL}(q \parallel p) = \frac{1}{2} \left( \text{tr}(\Sigma_p^{-1} \Sigma_q) + (\mu_p - \mu_q)^\top \Sigma_p^{-1} (\mu_p - \mu_q) - k + \ln \left( \frac{\det \Sigma_p}{\det \Sigma_q} \right) \right) \quad (3)$$

όπου  $k$  είναι η διάσταση του λανθάνοντος χώρου,  $\mu$  οι μέσες τιμές και  $\Sigma$  οι πίνακες διακύμανσης.

## 5 Αποτελέσματα

### 5.1 Standard Ρύθμιση

- **Device:** cpu
- **Loaded model from:** 'dreamer\_best.pth'

Η εκπαίδευση πραγματοποιήθηκε με την ακόλουθη διαμόρφωση:

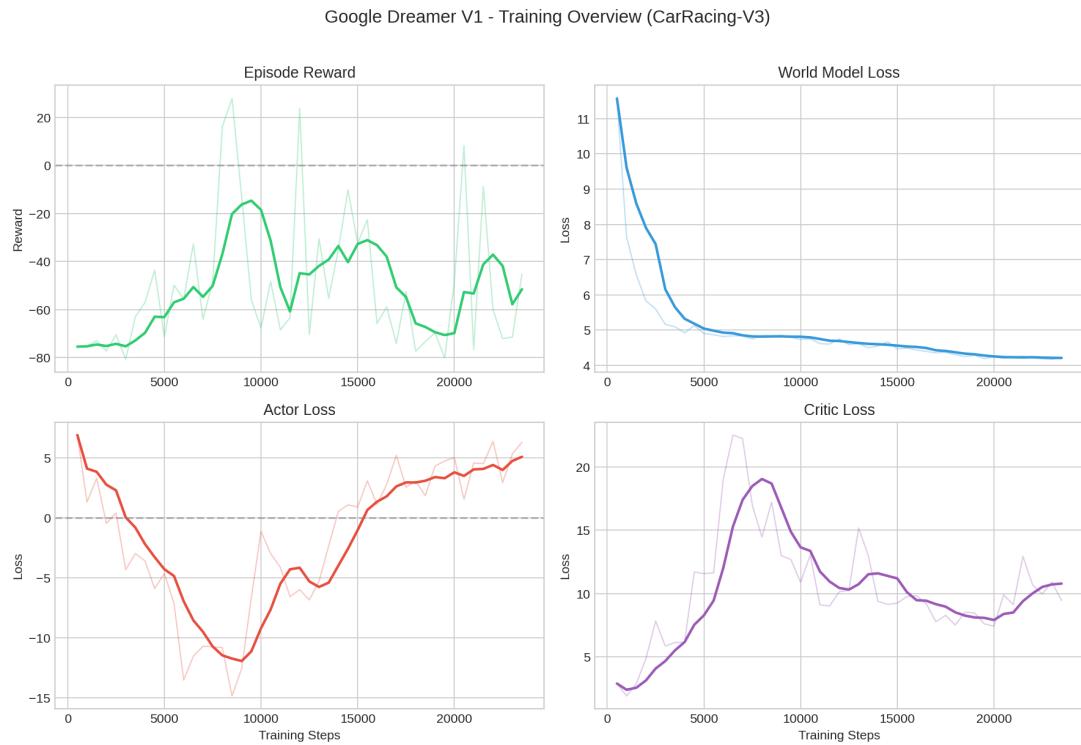
Παράμετρος	Τιμή
<i>Περιβάλλον</i>	
Environment	CarRacing-V3
Observation	64 × 64 grayscale
Action space	[steering, gas, brake] ∈ [−1, 1] <sup>3</sup>
Action repeat	2
<i>Αρχιτεκτονική</i>	
Stochastic dim (z)	30
Deterministic dim (h)	200
Hidden dim	300
Embedding dim	1024
<i>Εκπαίδευση</i>	
Total steps	500,000 (target) / 23,500 (actual)
Seed episodes	5
Batch size	50
Sequence length	50
Imagination horizon	15
<i>Hyperparameters</i>	
World model LR	$6 \times 10^{-4}$
Actor LR	$8 \times 10^{-5}$
Critic LR	$8 \times 10^{-5}$
Discount ( $\gamma$ )	0.99
Lambda ( $\lambda$ )	0.95
Free nats	3.0
Gradient clip	100
<i>Hardware</i>	
Device	CPU
Training time	~ 48 hours

**Πίνακας 1:** Παράμετροι εκπαίδευσης

### 5.2 Σύνοψη Εκπαίδευσης

Το μονέλο εκπαιδεύτηκε συνολικά για 47 επεισόδια, με βήματα που κυμαίνονται από 500 έως 23,500. Τα στατιστικά εκπαίδευσης δείχνουν σημαντική μείωση στο σφάλμα ανακατασκευής του μοντέλου, αν και το σήμα επιβράβευσης παρουσιάζει

υψηλή διακύμανση, κάτι που είναι τυπικό για την αρχική φάση της ενισχυτικής μάθησης σε σύνθετα περιβάλλοντα όπως το CarRacing-V3.



**Σχήμα 3:** Συνολική επισκόπηση της πορείας εκπαίδευσης (*Training Overview*).

Παρατηρήσαμε ένα θέμα στην εκπαίδευση του actor με το loss να αυξάνεται πολύ μετά από κάποια επεισόδια. Αυτό οφειλόταν σε μη σωστή υλοποίηση της εκπαίδευσης του actor όπου χρησιμοποιήσαμε detached actor states. Με την διόρθωση αυτού του σφάλματος, το μοντέλο κατάφερε να βελτιώσει σημαντικά την απόδοσή του. Ωστόσο λόγω απρόοπτης διακοπής της εκπαίδευσης δεν αποθηκευθηκαν τα logs αλλά μόνο το τελικό μοντέλο. Τεστάραμε το αρχικό και τελικό μοντέλο και παρουσιάζουμε τα αποτελέσματα παρακάτω.

### 5.3 Σύγκριση Επιδόσεων (Performance Comparison)

Metric	Initial Model	Optimized Model (Final)
Mean Reward	-42.07	<b>670.21</b>
Std Deviation	8.17	165.03
Min Reward	-53.18	372.13
Max Reward	-27.97	837.04

**Πίνακας 2:** Σύγκριση επιδόσεων μεταξύ αρχικού και τελικού μοντέλου. Με έντονο χρώμα ο κινούμενος μέσος όρος των τιμών και με αχνό χρώμα η πραγματική τιμή σε κάθε βήμα εκπαίδευσης.

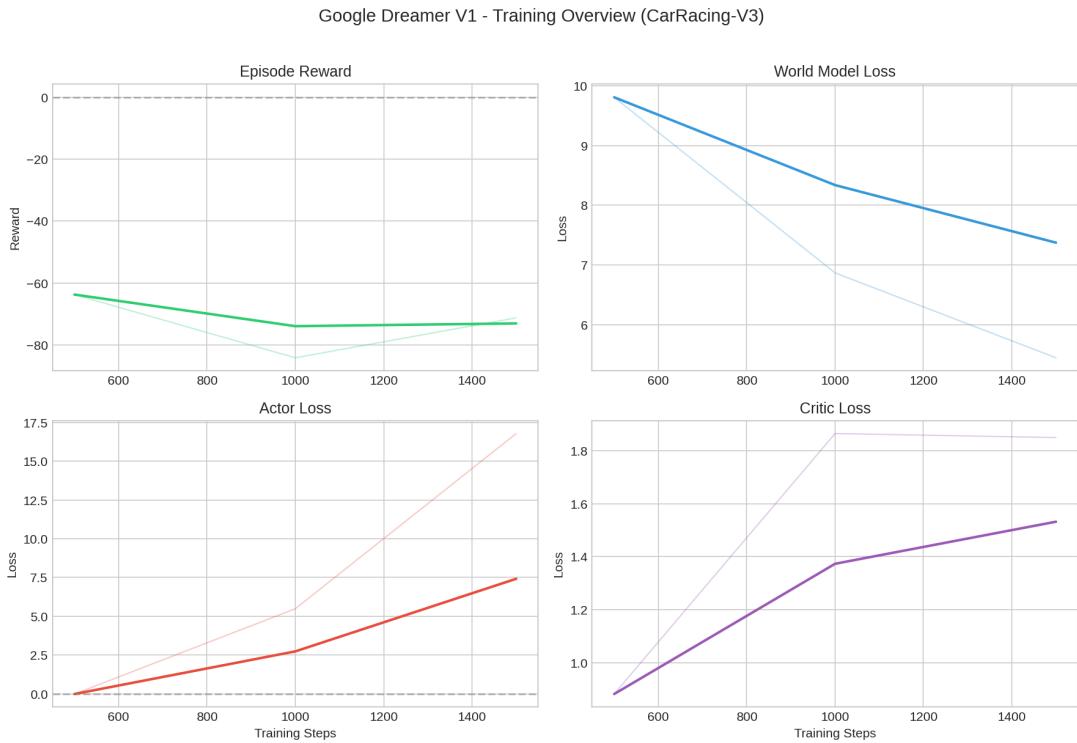
Αναλυτικά τα 5 επεισόδια της τελικής αξιολόγησης:

Episode	Steps	Reward
1	1000	837.04
2	1000	654.97
3	1000	678.17
4	1000	372.13
5	1000	808.75

**Πίνακας 3:** Αναλυτικά αποτελέσματα του βελτιστοποιημένου μοντέλου.

Το υψηλό μέσο reward (670.21) υποδεικνύει ότι ο agent έμαθε επιτυχώς να πλοηγείται στην πίστα, παραμένοντας εντός δρόμου και διατηρώντας ικανοποιητική ταχύτητα. Η μεγάλη απόκλιση (165.03) αντανακλά την ποικιλία στις επιδόσεις ανά επεισόδιο, πιθανώς λόγω διαφορετικών διαδρομών και συνθηκών στην πίστα. Αυτό είναι αναμενόμενο όταν η εκπαίδευση τερματίζεται μετά από λίγα επεισόδια, καθώς ο agent δεν έχει ακόμη γενικεύσει πλήρως την πολιτική του.

Προκειμένου να ξαναδημιουργήσουμε τις χαμένες καμπύλες επαναλάβαμε την διαδικασία της εκπαίδευσης η οποία διακόπηκε ξανά κατά την διάρκεια της νύχτας. Ωστόσο με επείγουσα αποθήκευση των logs καταφέραμε και σώσαμε την πρόοδο για τα 1500 πρώτα βήματα εκπαίδευσης. Αυτά προφανώς δεν είναι αρκετά βήματα σε σύγκριση με τα 5 εκατομύρια βήματα της google αλλά δίνουν μια εικόνα της προόδου.



**Σχήμα 4:** Συνολική επισκόπηση της πορείας εκπαίδευσης (*Training Overview*).

Γενικά δεν βλέπουμε σταθερή μείωση στα losses αφού στα στάδια της εξερεύνησης ο πράκτορας δοκιμάζει νέες ενέργειες που μπορεί να αυξήσουν το σφάλμα προσωρινά. Γι αυτό εξάλλου στα διαγράμματα έχουμε με έντονο χρώμα τον μέσο όρο των σφαλμάτων και με αχνό χρώμα την πραγματική τιμή σε κάθε βήμα εκπαίδευσης. Αν δούμε την πραγματική τιμή παρατηρούμε πιο απότομες διακυμάνσεις. Προφανώς με τόσα λίγα βήματα εκπαίδευσης η δειγματοληψία ανά 500 βήματα δεν είναι αρκετή για να έχουμε καλή εποπτεία.

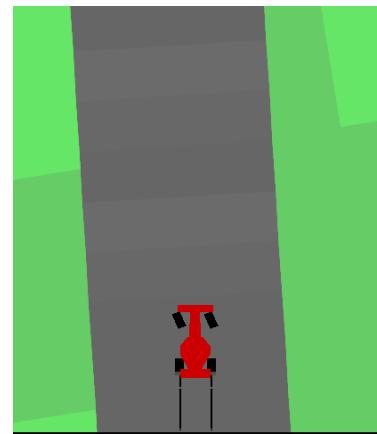
## 5.4 Οπτική Παρατήρηση σε διάφορα στάδια εκπαίδευσης

Παρακάτω δείχνουμε κάποιες εικόνες από το περιβάλλον προσομοίωσης με μία γρήγορη εκπαίδευση 5000 βημάτων. Η πρώτη μας δείχνει ότι το αυτοκίνητο παρόλο που ο δρόμος είναι ευθεία πάει να στρίψει. Στην δεύτερη έχει βγει εκτός πορείας.

Παρατηρούμε λοιπόν ότι το μοντέλο δεν έχει μάθει ακόμα να κρατιέται στην πορεία, δεν καταλαβαίνει καν τι χρειάζεται να κάνει. Αυτό είναι λογικό και αναμενόμενο αφού η μέθοδος του Dreamer απαιτεί πολλές ώρες εκπαίδευσης και μεγάλο αριθμό βημάτων δοκιμής και αποτυχίας μέχρι να φτάσει σε ικανοποιητικά αποτελέσματα.



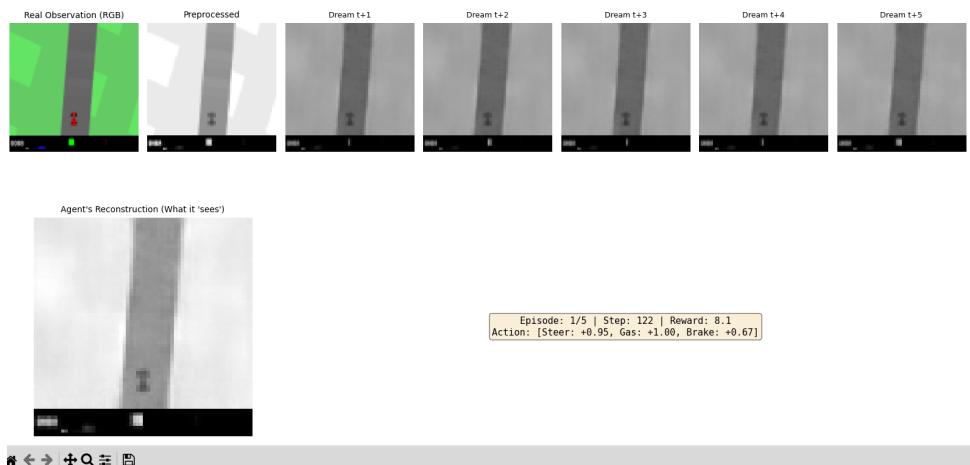
(α') Αυτοκίνητο που στρίβει σε ευθεία



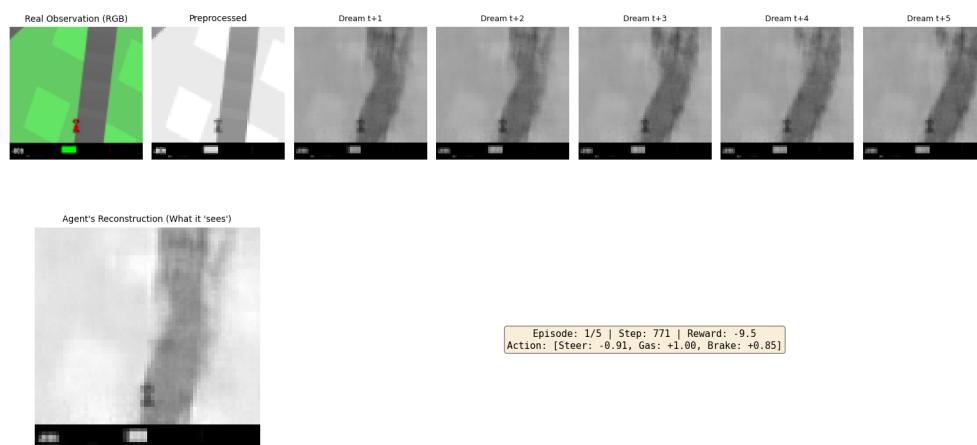
(β') Αυτοκίνητο εκτός πορείας

**Σχήμα 5: Εικόνες από την προσομοίωση**

Παρακάτω βλέπουμε πώς απεικονίζονται τα όνειρα κατά την διάρκεια του testing.



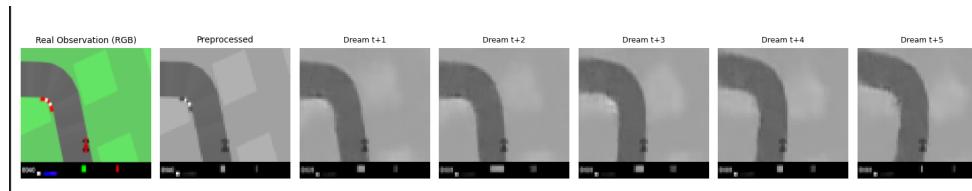
(α') Αυτοκίνητο που κινείται κανονικά σε ευθεία, 5 όνειρα και ανακατασκευή



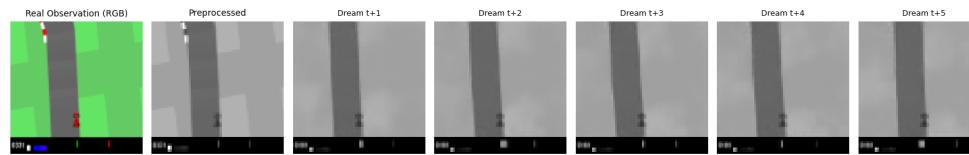
(β') Αυτοκίνητο οριακά εκτός πορείας σε ευθεία, 5 όνειρα και ανακατασκευή

### Σχήμα 6: Εικόνες από το testing με όνειρα και ανακατασκευές

Στη συνέχεια πραγματοποιήσαμε εκπαίδευση για 23500 βήματα και παρακάτω παρουσιάζουμε τα όνειρα (dreams) και τις ανακατασκευές (reconstructions) που παράγει το μοντέλο.



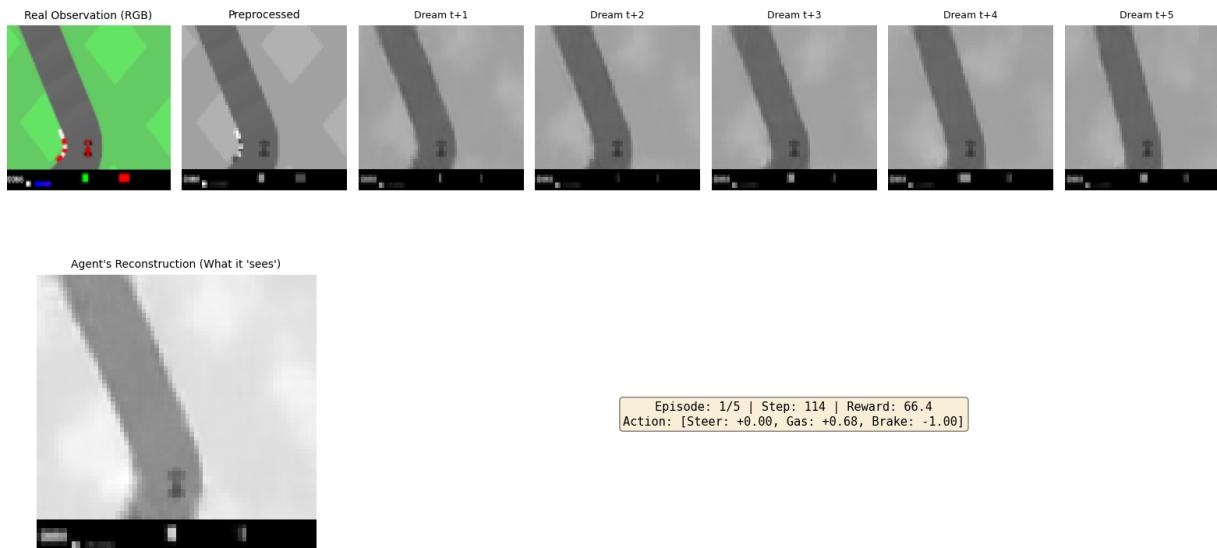
(α') Αυτοκίνητο εντός πορείας σε στροφή και 5 όνειρα



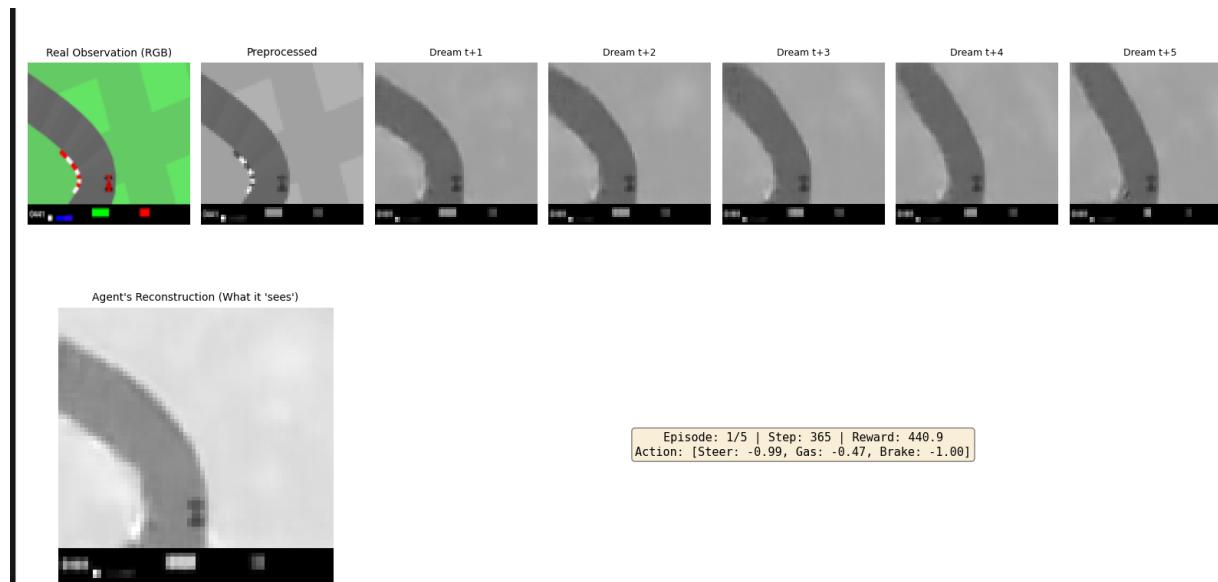
(β') Αυτοκίνητο οριακά εντός πορείας σε ευθεία και 5 όνειρα

### Σχήμα 7: Εικόνες από το testing με όνειρα

Παρατηρούμε ότι με 23000 βήματα εκπαίδευσης τα αποτελέσματα είναι πολύ καλύτερα αλλά κάποιες φορές πάει να βγει εκτός πορείας. Τέλος, παρακάτω παρουσιάζουμε όνειρα και ανακατασκευές από ένα καλύτερο run.



Σχήμα 8: Επιτυχημένη οδήγηση με όνειρα και ανακατασκευή



**Σχήμα 9:** Επιτυχημένη οδήγηση με όνειρα και ανακατασκευή (απότομη στροφή)

Παρατηρούμε ότι τα όνειρα διατηρούν βασικά χαρακτηριστικά της πραγματικής εικόνας, όπως η θέση του δρόμου και του αυτοκινήτου, αλλά ταυτόχρονα παρουσιάζουν παραμορφώσεις και απώλειες λεπτομερειών, όσο εξελισσεται ο δρόμος και ειδικά προς τις στροφές. Αυτή είναι η δουλειά του ονείρου να προβλέπει την μελλοντική κατάσταση βασισμένο στην τρέχουσα, χωρίς να έχει απόλυτη βεβαιότητα/ακριβεια.

Οι ανακατασκευές είναι αρκετά πιστές στις αρχικές εικόνες, υποδεικνύοντας ότι το VAE κατάφερε να μάθει μια αποτελεσματική αναπαράσταση του περιβάλλοντος τόσο για κωδικοποίηση όσο και για ανακατασκευή.

Ακόμα και σε κλειστές στροφές βρίσκει τον τρόπο να παραμείνει στην πορεία, δείχνοντας ότι ο πράκτορας έχει μάθει μια ικανοποιητική πολιτική οδήγησης.

## 6 Συμπεράσματα

Στην παρούσα εργασία υλοποιήσαμε τον αλγόριθμο Google Dreamer V1 από την αρχή και τον εφαρμόσαμε στο περιβάλλον CarRacing-V3 του Gymnasium. Η απόδοση του Google Dreamer V1 ανέδειξε τη σημαντική υπεροχή των Model-Based προσεγγίσεων σε περιβάλλοντα περιορισμένων πόρων.

Η χρήση του Λανθάνοντος Χώρου (Latent Space) επέτρεψε στον πράκτορα να γενικεύει αποτελεσματικά, μαθαίνοντας να οδηγεί ακόμη και όταν οι προβλέψεις του (όνειρα) γίνονταν ασαφείς σε βάθος χρόνου. Η δυνατότητα αυτή επιταχύνει την εκπαίδευση και μειώνει την ανάγκη για εκτεταμένη αλληλεπίδραση με το περιβάλλον. Η στρατηγική επιλογή της ασπρόμαυρης επεξεργασίας απέδειξε ότι η γεωμετρική πληροφορία επαρκεί για την επίλυση του CarRacing-V3, μειώνοντας δραστικά το υπολογιστικό κόστος. Τέλος, διαπιστώθηκε πρακτικά ότι η σωστή διάδοση των gradients μέσω των φανταστικών τροχιών (backpropagation through time) είναι απαραίτητη για την επιτυχή εκπαίδευση του Actor, καθώς χωρίς αυτήν ο πράκτορας αδυνατεί να συνδέσει τις ενέργειές του με τα μελλοντικά αποτελέσματα.

Η αλληλεξάρτηση των διαφόρων συνιστωσών του Dreamer (VAE, RSSM, Actor, Critic) κατά την εκπαίδευση είναι συγχρόνως προτέρημα και πρόκληση, γεγονός που κάνει τον Google Dreamer V1 έναν ισχυρό και έξυπνο αλγόριθμο για την επίλυση σύνθετων προβλημάτων ενισχυτικής μάθησης.

Θα ήταν ενδιαφέρον να τον εκπαιδεύσουμε και σε άλλα περιβάλλοντα όπως το "DOOM" ή το "Atari" ωστόσο λόγω των περιορισμένων υπολογιστικών πόρων δεν ήταν εφικτό να το πραγματοποιήσουμε στο πλαίσιο αυτής της εργασίας.

## Αναφορές

---

- [1] Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2020). *Dream to Control: Learning Behaviors by Latent Imagination*. International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1912.01603>
- [2] Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). *Learning Latent Dynamics for Planning from Pixels*. International Conference on Machine Learning (ICML). <https://arxiv.org/abs/1811.04551>
- [3] Ha, D., & Schmidhuber, J. (2018). *World Models*. <https://arxiv.org/abs/1803.10122>
- [4] Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1312.6114>
- [5] Wikipedia contributors. (2024). *Variational autoencoder*. Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Variational\\_autoencoder](https://en.wikipedia.org/wiki/Variational_autoencoder)
- [6] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. EMNLP. <https://arxiv.org/abs/1406.1078>
- [7] Farama Foundation. (2023). *Gymnasium: A Standard API for Reinforcement Learning*. <https://gymnasium.farama.org/>