Selección Dinámica de Datos Visuales para Modelos del Mundo

Bruno Corcuera[1], Carlos Eiras-Franco[1,2], Brais Cancela[1,2]

[1] Universidade da Coruña, A Coruña, España
[2] CITIC Research Center, Universidade da Coruña, A Coruña, España
bruno.sanchez1@udc.es

Resumen Los modelos del mundo son sistemas que aprenden una representación interna y compacta de un entorno para predecir su evolución y tomar decisiones. En tareas visuales, algunas regiones de cada imagen pueden ser redundantes: por ejemplo, fondos vacíos o texturas repetidas. Proponemos Selección Dinámica de Datos (DDS, por sus siglas en inglés), un módulo que aprende sin etiquetas qué información visual conviene conservar en cada observación antes de codificarla. DDS mantiene la posición original de los elementos seleccionados y limita de forma explícita la cantidad de información que recibe el codificador. A diferencia de los métodos que ocultan partes de una imagen solo durante el preentrenamiento, DDS utiliza la entrada seleccionada tanto durante el entrenamiento como durante la inferencia. Evaluamos el método en modelos del mundo mediante la reconstrucción de imágenes, generación de vídeo y control. En una arquitectura modular basada en un autocodificador variacional (VAE), DDS reduce el error de reconstrucción usando una fracción pequeña de la entrada, mejora la coherencia temporal de las secuencias generadas y obtiene un mayor rendimiento del controlador bajo el mismo protocolo. Además, su integración en DreamerV3 aumenta el retorno medio en CarRacing, Crafter y MsPacman. Los experimentos que estudian por separado los componentes del método indican que la selección espacial aprendida contribuye a la mejora y que DDS resulta especialmente útil cuando las observaciones contienen información visual redundante.

Keywords: Selección Dinámica de Datos · Selección de información visual · Modelos del mundo · Entradas visuales dispersas · Control desde imágenes

1. Introducción

Un modelo del mundo es un sistema que aprende una representación interna simplificada de un entorno a partir de sus observaciones. En lugar de conservar cada detalle, resume información útil para reconstruir lo que el agente observa, predecir qué puede ocurrir a continuación y elegir acciones [8,12]. Cuando las observaciones son imágenes, el modelo debe condensar una gran cantidad de información. Sin embargo, no todos los píxeles son igual de relevantes para estos

objetivos: una imagen puede contener fondos vacíos, texturas repetidas o detalles poco relacionados con la dinámica del entorno.

Habitualmente, un codificador procesa la imagen completa y la comprime en un vector compacto. En este trabajo estudiamos una opción complementaria: seleccionar primero la información visual relevante y codificar después únicamente esa entrada filtrada. Esta idea está relacionada con la selección de características, que conserva parte de la entrada original en lugar de sustituirla por variables transformadas [33,7]. Estudiamos si la representación puede mejorar cuando el codificador recibe menos información, seleccionada según su relevancia para el objetivo de aprendizaje.

Una selección fija no es suficiente para este problema. En una secuencia visual, los objetos se mueven y las regiones importantes cambian de una imagen a otra. Por ello, necesitamos adaptar la selección a cada observación. Algunos métodos previos realizan selecciones dependientes de la entrada, aunque suelen centrarse en problemas supervisados o datos tabulares [6,32,1]. También existen métodos que ocultan regiones de una imagen para aprender representaciones visuales mediante máscaras [3,15,31,30]. En general, esas regiones se ocultan para crear una tarea de preentrenamiento, no para definir la entrada que usará finalmente el modelo.

Proponemos Selección Dinámica de Datos (DDS), un módulo que aprende sin etiquetas qué elementos de cada imagen debe conservar. DDS se coloca antes del codificador visual y predice la importancia de cada dato de entrada de manera aislada. Para su funcionamiento, construye una máscara que conserva como máximo los M elementos más relevantes y oculta el resto. La entrada resultante mantiene la forma y la organización espacial originales de la imagen. Así, el codificador puede procesarla directamente y el parámetro M permite controlar cuánta información visual recibe.

Usamos modelos del mundo como marco de evaluación porque permiten analizar más que la reconstrucción de imágenes. Una representación visual adecuada también debe permitir generar secuencias coherentes y conservar información útil para actuar. Comparamos DDS con un autocodificador variacional convencional

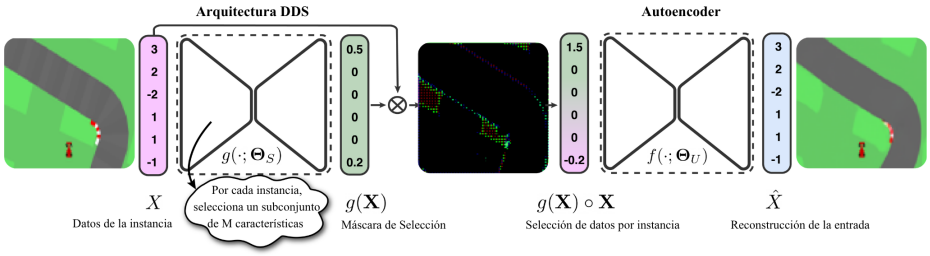


Figura 1. Selección Dinámica de Datos. DDS se coloca antes del codificador visual, estima la importancia de los datos de entrada y genera una máscara distinta para cada observación sin alterar la disposición espacial original.

(VAE, por sus siglas en inglés) y con una referencia MAE+VAE, que combina un autocodificador enmascarado (MAE, por sus siglas en inglés) y un VAE. Además, incluimos un control sin selección para distinguir la mejora asociada al módulo auxiliar de la aportada por la selección aprendida. Finalmente, integramos DDS en DreamerV3 para estudiar si el método también funciona en un agente moderno entrenado conjuntamente.

Las contribuciones principales son dos. Primero, presentamos DDS, un método sin etiquetas que selecciona para cada imagen la información visual que llega al codificador. Segundo, mostramos que esta selección puede mejorar la reconstrucción, la generación de secuencias y el control en modelos del mundo. También analizamos sus límites mediante experimentos que estudian sus componentes por separado y comparan distintos niveles de selección.

2. Trabajo relacionado

2.1. Selección fija y selección dependiente de la entrada

Seleccionar características significa conservar parte de las variables observadas. Muchos métodos clásicos y diferenciables aprenden una selección fija para todo el conjunto de datos [7,2,25,5]. Esta opción es adecuada cuando la relevancia de las variables puede aproximarse de forma global. En una imagen, sin embargo, los objetos, los bordes y las señales de movimiento pueden aparecer en posiciones distintas.

Otros métodos adaptan la selección a cada entrada. L2X, INVASE y TabNet son ejemplos de esta idea, principalmente en problemas supervisados o con datos tabulares [6,32,1]. DDS sigue el mismo principio, pero selecciona información espacial en imágenes sin etiquetas y aprende mediante la función de pérdida del modelo al que precede. También guarda relación con métodos de atención y transformación espacial: algunos recorren regiones de forma secuencial y otros aprenden transformaciones diferenciables dentro de redes convolucionales [24,18]. DDS realiza una operación más directa: oculta píxeles de la entrada original y entrega al codificador una imagen enmascarada con la misma disposición espacial.

2.2. Máscaras y modelos del mundo

Otra línea de trabajo oculta partes de una imagen durante el preentrenamiento para que el modelo recupere la información ausente. BEiT, MAE, SimMIM y MaskFeat predicen distintos tipos de unidades ocultas, como códigos visuales, píxeles o características visuales [3,15,31,30]. En estos métodos, la máscara define una tarea de aprendizaje. DDS tiene otro propósito: la entrada seleccionada es la que usa el codificador tanto durante el entrenamiento como durante la inferencia. Masked World Models también enmascara información al aprender modelos del mundo [27], pero separa ese aprendizaje visual de la dinámica latente. DDS aprende la selección de cada entrada mediante la función de pérdida del modelo al que precede.

Los modelos del mundo aprenden estados internos compactos para predecir, generar secuencias y actuar. La formulación modular original separa la visión, la memoria recurrente y el controlador [8]. PlaNet y Dreamer incorporan dinámica latente estocástica para planificar o aprender políticas directamente desde píxeles. DreamerV3 y TD-MPC2 extienden este principio a escenarios de control más amplios [11,10,12,13]. DDS aborda una cuestión complementaria: si el estado visual puede mejorar al seleccionar explícitamente la información que llega al codificador.

3. Selección Dinámica de Datos

DDS se coloca antes de una arquitectura visual y sustituye la entrada original por una observación parcialmente enmascarada. Primero predice la importancia de cada dato de entrada, de manera aislada; por ejemplo, de cada píxel. Para su funcionamiento, construye una máscara que conserva como máximo las M unidades más importantes y oculta el resto. La selección cambia para cada observación y mantiene la disposición espacial de la imagen. El método no requiere etiquetas: aprende mediante la misma pérdida de reconstrucción, representación, predicción o modelado del mundo con la que se entrena el módulo posterior.

Para expresar esta idea formalmente, sea $\mathbf{X} \in \mathbb{R}^{N \times F}$ un lote de N observaciones, donde F es el número de unidades seleccionables tras aplanar cada entrada. El aplanamiento solo simplifica la notación: en una imagen, la máscara conserva la rejilla original. DDS busca para cada observación un subconjunto de como máximo M unidades:

$$\begin{aligned} \underset{\boldsymbol{\Theta}_S, \boldsymbol{\Theta}_U}{\text{minimizar}} \quad & \mathcal{L}(f(g(\mathbf{X}; \boldsymbol{\Theta}_S) \odot \mathbf{X}; \boldsymbol{\Theta}_U)) \\ \text{sujeto a} \quad & g(\mathbf{X}; \boldsymbol{\Theta}_S) \in [0,1]^{N \times F}, \\ & \left\| g(\mathbf{X}; \boldsymbol{\Theta}_S)^{(i)} \right\|_0 \leq M, \quad \forall i \in \{1, \ldots, N\}. \end{aligned} \tag{1}$$

Aquí $g(\cdot; \boldsymbol{\Theta}_S)$ es el selector, $f(\cdot; \boldsymbol{\Theta}_U)$ el modelo posterior y $\mathcal{L}$ su pérdida. A diferencia de una selección fija para todo el conjunto de datos, DDS produce una máscara nueva para cada observación. El selector estima la importancia de las F unidades y el límite M se aplica después.

Elegir las unidades que se conservan es una operación discreta, pero el selector debe entrenarse por descenso de gradiente. Para compatibilizar ambos requisitos, DDS combina una compuerta continua con una máscara binaria. Sean $\tilde{g}(\mathbf{X}; \boldsymbol{\Theta}_S)$ los valores sin normalizar del selector y $\boldsymbol{\Gamma}_M \in \{0,1\}^{N \times F}$ una matriz binaria que conserva las unidades asociadas a los M valores mayores de cada observación. La entrada efectiva es

$$\mathbf{X}_{\text{DDS}} = \tau(\tilde{g}(\mathbf{X}; \boldsymbol{\Theta}_S) + \delta) \odot \boldsymbol{\Gamma}_M \odot \mathbf{X}. \tag{2}$$

La compuerta continua $\tau(\cdot)$ pondera las unidades retenidas, mientras que $\boldsymbol{\Gamma}_M$ impone el límite durante la propagación hacia delante. Las unidades no seleccionadas quedan ocultas para el codificador.

Para definir la compuerta continua, DDS usa una función recortada de tipo *hard-concrete*, basada en relajaciones continuas de variables discretas [23,19,21]:

$$\tau(x) = \text{mín}\left(1, \text{máx}\left(0, \sigma\left(\frac{x}{\beta}\right)(\zeta - \gamma) + \gamma\right)\right), \tag{3}$$

donde σ es la sigmoide, β la temperatura y γ y ζ los límites del intervalo extendido. En la Ec. 2, δ desplaza los valores del selector antes de aplicar la compuerta. Como $\boldsymbol{\Gamma}_M$ ya impone el número máximo de unidades activas, no se añade una penalización L_0. Usamos por defecto $\zeta = 1$, $\gamma = 0$, $\delta = 1$, $\beta = \frac{2}{3}$ y una variante estocástica controlada por $\kappa = 0.1$ durante el entrenamiento. En evaluación, $\kappa = 0$ hace determinista la máscara.

Cuando $M \ll F$, aplicar el límite desde el inicio puede ocultar casi toda la información útil antes de que el selector aprenda a identificarla. Para facilitar el aprendizaje inicial, DDS relaja ocasionalmente la restricción de cardinalidad durante el entrenamiento. Para cada observación i en el paso t,

$$M_t^{(i)} = \begin{cases} M, & p_t^{(i)} > \epsilon, \quad p_t^{(i)} \sim \mathcal{U}(0,1), \\ F, & \text{en otro caso.} \end{cases} \tag{4}$$

Con probabilidad ϵ, el límite se amplía temporalmente de M a F, de modo que ninguna unidad queda excluida por la máscara binaria. En inferencia, el límite siempre es M.

Implementamos $g(\cdot; \boldsymbol{\Theta}_S)$ con una red convolucional ligera de estilo U-Net [26], que combina detalle local y contexto a varias escalas sin perder alineamiento espacial. En los experimentos, DDS se entrena primero junto a una arquitectura convolucional que reconstruye la entrada a partir de $\mathbf{X}_{\text{DDS}}$. Después se congela esta arquitectura y se entrena un VAE compacto sobre su representación intermedia. La pérdida del VAE combina reconstrucción capa a capa a través de la arquitectura congelada y un término KL con la técnica *free-bits* ($\lambda = 0.02$), que evita incentivar la reducción del término KL por debajo de un margen [20]. No se usa una pérdida perceptual procedente de una red externa. Esta separación permite evaluar si la selección visual mejora el estado latente sin rediseñar por completo el modelo del mundo.

4. Evaluación en modelos del mundo

4.1. Protocolo experimental

La evaluación principal mide si DDS produce estados visuales útiles para reconstruir imágenes, generar vídeo y controlar un agente. Los experimentos modulares siguen el protocolo de datos, dimensión latente, memoria recurrente y evaluación del controlador del modelo del mundo original [8]. En todas las variantes, el módulo de visión transforma cada observación en un vector latente $\mathbf{z} \in \mathbb{R}^{32}$. En DDS+VAE, DDS no reemplaza a este estado latente: primero genera una observación enmascarada y después el codificador del VAE la comprime en $\mathbf{z}$.

Este vector, no la máscara, es la representación que recibe la memoria recurrente y que se usa después para control. Para `CarRacing-v3` y `SuperMarioBros-v0`, se recopilaron 10 millones de observaciones por entorno (10 000 secuencias de 1 000 pasos cada una). La reconstrucción se evalúa sobre 100 000 fotogramas reservados para prueba.

Comparamos el VAE original, una referencia MAE+VAE y DDS+VAE. Como DDS+VAE añade una arquitectura convolucional auxiliar alrededor del VAE compacto, incluimos un control U-Net que deja pasar la imagen completa. Así distinguimos el efecto de esta arquitectura del efecto de la selección aprendida.

4.2. Reconstrucción con distintos límites de selección

Medimos la reconstrucción mediante el error cuadrático medio (MSE). Un valor menor indica que la imagen reconstruida se aproxima más a la original.

Tabla 1. MSE de reconstrucción en 100 000 fotogramas reservados de `CarRacing-v3`. Las columnas indican la fracción de entrada conservada. Menor es mejor.

Módulo visual	100 %	32 %	16 %	8 %	4 %	2 %
VAE	0.00165	–	–	–	–	–
MAE+VAE	0.00126	0.00220	0.00263	0.00317	0.00388	0.00475
U-Net completa + VAE	0.00121	–	–	–	–	–
DDS + VAE	–	0.00051	0.00047	0.00042	**0.00039**	0.00134

La reconstrucción de un solo fotograma permite evaluar el codificador visual de forma aislada antes de analizar secuencias y control. La Tabla 1 muestra que DDS+VAE obtiene el menor error al conservar un 4 % de la entrada: alcanza un MSE de 0.00039, frente a 0.00165 en el VAE estándar y 0.00121 en el control U-Net que recibe la imagen completa. Esta diferencia indica que añadir una arquitectura convolucional auxiliar no basta para explicar la mejora. La selección aprendida contribuye de forma sustancial al resultado.

4.3. Calidad de secuencias generadas

Un estado visual de un modelo del mundo debe permitir generar trayectorias coherentes tanto en apariencia como en evolución temporal. Generamos 100 secuencias de 1 000 fotogramas por entorno y las comparamos con trayectorias reales mediante Fréchet Inception Distance (FID) y Fréchet Video Distance (FVD) [16,29]. FID compara la distribución visual de los fotogramas. FVD tiene en cuenta además la estructura temporal de las secuencias. En ambos casos, menor es mejor.

La Tabla 2 muestra que DDS+VAE mejora simultáneamente FID y FVD en `CarRacing-v3`. En `SuperMarioBros-v0`, MAE+VAE alcanza un FID ligeramente menor, pero DDS+VAE obtiene el mejor FVD.

Tabla 2. Calidad de las secuencias generadas. DDS+VAE conserva un 4 % de la entrada. Menor FID/FVD es mejor.

Entorno	Métrica	VAE	MAE+VAE	DDS+VAE
`CarRacing-v3`	FID ↓	59.46	54.84	**25.35**
`CarRacing-v3`	FVD ↓	239	312	**176**
`SuperMarioBros-v0`	FID ↓	64.06	**60.21**	61.08
`SuperMarioBros-v0`	FVD ↓	412	465	**338**

4.4. Control y transferencia a DreamerV3

Para comprobar si el estado seleccionado sigue siendo útil para actuar, entrenamos controladores a partir de los estados latentes producidos por VAE y DDS+VAE bajo el mismo protocolo CMA-ES [14] y evaluamos 100 episodios. El VAE original obtiene un retorno medio de 734.96 ± 162.75, mientras que DDS+VAE alcanza 818.58 ± 147.05. La diferencia es menor que en reconstrucción o generación, pero indica que DDS no solo mejora las reconstrucciones: el estado conserva información útil para elegir acciones.

También integramos DDS en DreamerV3, un agente moderno basado en dinámica latente y entrenado conjuntamente [12]. DDS se inserta en la arquitectura de codificación visual sin modificar el resto del protocolo de entrenamiento. Evaluamos `CarRacing-v3`, `Crafter` como entorno de supervivencia de mundo abierto y `MsPacman` [28,9,4,22].

Tabla 3. Transferencia a DreamerV3. Se reporta retorno medio ± desviación típica en el agente final evaluado.

Entorno	DreamerV3	DreamerV3-DDS	Cambio
`CarRacing-v3`	714.96 ± 232.63	$\mathbf{832.15 \pm 111.44}$	+16.4 %
`Crafter`	7.12 ± 1.07	$\mathbf{8.12 \pm 1.21}$	+14.0 %
`MsPacman`	1305.33 ± 620.89	$\mathbf{1717.00 \pm 472.37}$	+31.5 %

La Tabla 3 muestra que DreamerV3-DDS obtiene un mayor retorno medio en los tres entornos. Estos resultados indican que seleccionar información visual antes de comprimirla también puede ser útil fuera de la arquitectura modular basada en VAE.

4.5. Coste computacional

DDS limita la información que recibe el codificador, pero el selector también tiene un coste. En los módulos visuales modulares, DDS+VAE usa menos parámetros entrenables que VAE y MAE+VAE (4.04 millones frente a 4.35 y 6.69

millones), aunque aumenta el tiempo de entrenamiento por el cálculo de la selección y el procedimiento en dos etapas: 175 minutos frente a 85 minutos para VAE y 100 minutos para MAE+VAE. En DreamerV3 sobre `CarRacing-v3`, 200 000 iteraciones tardan 10.74 horas en DreamerV3 y 11.98 horas en DreamerV3-DDS, un incremento del 11.5 %. Por tanto, DDS controla cuánta información visual llega al codificador, pero no debe interpretarse como una técnica general de aceleración.

5. Análisis por componentes y generalización

Las máscaras permiten inspeccionar qué información se presenta al codificador, aunque no deben interpretarse como explicaciones causales. La Figura 2 muestra que, al aumentar la fracción retenida en `CarRacing-v3`, la selección se concentra primero en bordes de pista y regiones próximas al vehículo, y después incorpora un contexto más amplio.

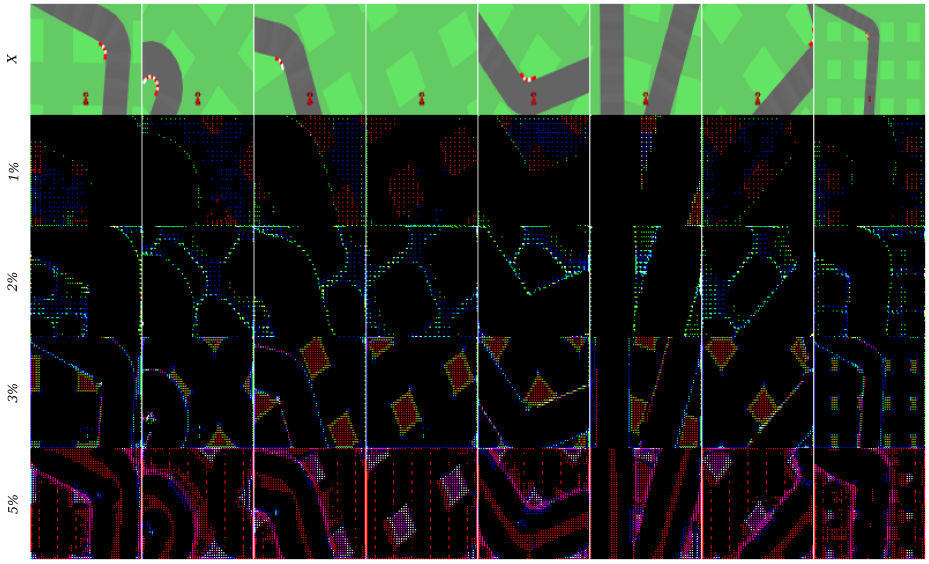


Figura 2. Máscaras DDS en `CarRacing-v3` con distintas fracciones de entrada retenida. La visualización ilustra cómo una selección pequeña se concentra en bordes y regiones espacialmente estructuradas.

Los experimentos por componentes indican que DDS se beneficia tanto de una arquitectura que conserva el alineamiento espacial como de mantener las M unidades con mayor importancia estimada. En `CarRacing-v3`, eliminar las conexiones residuales del selector empeora el MSE para todos los límites estudiados. Las diferencias entre compuertas son menores: la variante clásica *binary-concrete*

obtiene el mejor resultado al 4 % (0.00039), mientras que la configuración completa de DDS queda muy cerca con ese límite (0.00041) y es la mejor para $M = 2\,048$. Además, el resultado principal de DDS al 4 % obtiene un error muy inferior al del control U-Net que recibe la imagen completa. Por tanto, la capacidad de la arquitectura convolucional no basta para explicar la mejora.

El límite M es el parámetro principal de diseño. En un estudio de 48 configuraciones sobre reconstrucción de `Tiny ImageNet`, variando la temperatura β, la estocasticidad κ y la probabilidad ϵ de relajar la restricción de cardinalidad, el MSE medio fue 2.6×10^{-4} con desviación 1.7×10^{-4}. La configuración por defecto se situó entre las de menor error. En este escenario, los parámetros secundarios no requieren un ajuste minucioso, mientras que M sí debe elegirse según la redundancia visual y la fracción de información que se desea conservar.

Esa dependencia se observa también fuera de los modelos generativos. En clasificación, al incorporar DDS a ProPos [17], la variante con selección queda ligeramente por debajo de la versión con entrada completa en `CIFAR-10` cuando conserva una fracción muy pequeña de la entrada. Este resultado es coherente con imágenes compactas donde muchos píxeles pueden contener información útil. En conjuntos de mayor resolución o con más fondo, como `ImageNet-10`, `ImageNet-Dogs` y `Tiny ImageNet`, DDS mejora las métricas de agrupamiento usando menos unidades de entrada. La conclusión no es que DDS sea universalmente superior, sino que resulta útil cuando existe información visual prescindible y la función de pérdida proporciona una señal fiable para el selector.

6. Discusión y conclusiones

Los resultados apoyan una interpretación acotada de DDS: adaptar la selección espacial a cada observación ayuda cuando un codificador visual debe formar un estado compacto a partir de imágenes con información redundante o secundaria para el objetivo. En el modelo del mundo modular, DDS mejora la reconstrucción, la calidad de las secuencias generadas y el control. La comparación con una U-Net que recibe la imagen completa indica que la arquitectura auxiliar no basta para explicar la mejora. En DreamerV3, DDS mantiene la misma idea dentro de un agente entrenado conjuntamente y obtiene un mayor retorno medio en los tres entornos evaluados.

DDS se diferencia del aprendizaje autosupervisado mediante ocultación porque la máscara no es solo una alteración artificial utilizada durante el entrenamiento. La observación seleccionada es la entrada que procesa el codificador durante la inferencia, por lo que M controla explícitamente cuánta información contribuye al estado latente. Esta propiedad también define sus límites. Un valor de M demasiado pequeño puede eliminar señales útiles, especialmente en imágenes compactas; además, el selector añade coste de entrenamiento. Las máscaras DDS reflejan la información seleccionada para un objetivo, no explicaciones causales, y pueden capturar correlaciones espurias o atajos del conjunto de datos. No se ha evaluado explícitamente la robustez de DDS ante cambios en la distribución del entorno, por lo que este análisis queda como trabajo futuro.

En resumen, DDS ofrece un mecanismo práctico para aprender representaciones visuales a partir de entradas dispersas mediante selección explícita de información antes de la compresión. Sus mejoras en reconstrucción, generación y control sugieren que seleccionar la información presentada al codificador puede ser un componente relevante junto al diseño del propio cuello de botella latente. El trabajo futuro debería estudiar la selección automática del límite, selectores más ligeros e integraciones con modelos fundacionales o con representaciones visuales formadas por unidades discretas.

Agradecimientos Este trabajo ha sido financiado parcialmente por el MICIU/AEI/10.13039/501100011033 y el FEDER/UE mediante las ayudas PID2023-147404OB-I00 y PID2021-128045OA-I00; por el Ministerio para la Transformación Digital y de la Función Pública y NextGenerationEU/PRTR mediante la ayuda TSI-100925-2023-1; por la Xunta de Galicia mediante la ayuda ED431C 2022/44; y por la Consellería de Cultura, Educación e Universidade, en el marco del Programa FEDER Galicia 2021–2027, y la Secretaría Xeral de Universidades mediante la ayuda ED431G 2023/01.

Declaración de intereses Los autores declaran que no tienen intereses financieros ni relaciones personales conocidas que pudieran influir en el trabajo presentado.

Referencias

1. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 6679–6687 (2021)
2. Balın, M.F., Abid, A., Zou, J.: Concrete autoencoders: Differentiable feature selection and reconstruction. In: International conference on machine learning. pp. 444–453. PMLR (2019)
3. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022)
4. Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M.: The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research **47**, 253–279 (2013)
5. Cancela, B., Bolon-Canedo, V., Alonso-Betanzos, A.: E2e-fs: An end-to-end feature selection method for neural networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **45**(07), 8311–8323 (2023)
6. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: An information-theoretic perspective on model interpretation. In: International conference on machine learning. pp. 883–892. PMLR (2018)
7. Dhal, P., Azad, C.: A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence pp. 1–39 (2022)
8. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
9. Hafner, D.: Benchmarking the spectrum of agent capabilities. In: International Conference on Learning Representations (2022)
10. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)

11. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: International conference on machine learning. pp. 2555–2565. PMLR (2019)
12. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse control tasks through world models. Nature (2025). https://doi.org/10.1038/s41586-025-08744-2
13. Hansen, N., Su, H., Wang, X.: Td-mpc2: Scalable, robust world models for continuous control. In: International Conference on Learning Representations. vol. 2024, pp. 47376–47405 (2024)
14. Hansen, N.: The CMA evolution strategy: A tutorial. CoRR **abs/1604.00772** (2016), http://arxiv.org/abs/1604.00772
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (June 2022)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR **abs/1706.08500** (2017), http://arxiv.org/abs/1706.08500
17. Huang, Z., Chen, J., Zhang, J., Shan, H.: Learning representation for clustering via prototype scattering and positive sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
18. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025 (2015)
19. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017)
20. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improving variational inference with inverse autoregressive flow. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
21. Louizos, C., Welling, M., Kingma, D.P.: Learning sparse neural networks through l_0 regularization. In: International Conference on Learning Representations. pp. 1–13 (2018)
22. Machado, M.C., Bellemare, M.G., Talvitie, E., Veness, J., Hausknecht, M., Bowling, M.: Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. Journal of Artificial Intelligence Research **61**, 523–562 (2018)
23. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: International Conference on Learning Representations (2017)
24. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems. pp. 2204–2212 (2014)
25. Roffo, G., Melzi, S., Castellani, U., Vinciarelli, A., Cristani, M.: Infinite feature selection: a graph-based feature filtering approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4396–4410 (2020)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
27. Seo, Y., Hafner, D., Liu, H., Liu, F., James, S., Lee, K., Abbeel, P.: Masked world models for visual control. In: Conference on Robot Learning. pp. 1332–1344. PMLR (2023)

28. Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J.J., Tan, H., Younis, O.G.: Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032 (2024)
29. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. CoRR **abs/1812.01717** (2018), http://arxiv.org/abs/1812.01717
30. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
31. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: SimMIM: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
32. Yoon, J., Jordon, J., van der Schaar, M.: Invase: Instance-wise variable selection using neural networks. In: International Conference on Learning Representations. pp. 1–11 (2018)
33. Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., Saeed, J.: A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends **1**(2), 56–70 (2020)