A. Appendix / Supplemental material

A.1. Synthetic Dataset Dimensionality

Each experiment uses an image size of (50×50) pixels, and the simulated dynamics is the same for the three datasets (Eq. 2) normalized with respect to the image size and maximum pixel intensity. Each dataset consists of 500 training samples with 20 frames, with a final numerical dimensionality of $(samples \times frames \times \#channels \times width \times height) = (500 \times 20 \times 1 \times 50 \times 50)$.

A.2. Training Hyperparameters

The experiments and baselines were executed on an NVIDIA 3080 GPU. For our model, implemented in PyTorch, the encoder was trained using the Adam optimizer [25] with a learning rate of $1 \times e^{-2}$ and the Kaming weight initialization for MLP layers.

A.3. Simulation Details

Here we discuss the details of the dynamics simulation of the experiments in section 5.1. The equation 2 represents a harmonic oscillator with the closed solution:

$$z(t) = Ae^{-\zeta t} cos(\omega t + \phi). \tag{18}$$

Where $\omega = 2$ is the frequency we used for simulation and $\zeta = 0.04$ the damping factor. This parameter relates to γ as follows:

$$\gamma_0 = \omega^2 + \zeta^2 = 4.0016. \tag{19}$$

$$\gamma_1 = 2\zeta = 0.08. \tag{20}$$

A.4. Ablation study

We present an ablation study to show the effect of the KL-divergence (KLD) term in our loss function; we used the intensity experiment presented in the synthetic experiments section. Table 3 shows the comparison of learned parameters of the physical equations along with their expected, ground truth (GT) values. In addition, Figure 9 shows the convergence discussed in the methods section, where a shortcut for the model to optimize the mean squared error (MSE) in the latent space is to converge always to the mean value of the dynamic variable z.

Parameter	MSE+KLD	MSE	**GT**
γ_0	3.99	5.7	4
γ_1	0.08	6.6	0.08

Table 3. Ablation comparison of the two-term losses. The model relaying only MSE cannot learn the expected values, while the KLD term allows it to get a proper estimation.

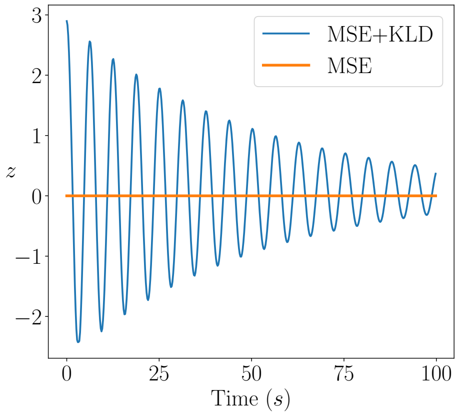


Figure 9. Comparison of the model trained with the MSE loss only (orange) and with our additional KLD loss term. The model relying just on MSE converges to the mean value of the dynamics, while our term guarantees that the model has the expected diversity.

A.5. Baseline Dataset

In this section, we describe the dataset used to compare the baselines. Both baselines were tested on the dataset published by [20] and also used in [17]. The equation of motion used for both systems is Eq. 21

$$\vec{F}_{ij} = -k(\vec{p}_i - \vec{p}_j) - l\frac{\vec{p}_i - \vec{p}_j}{|\vec{p}_i - \vec{p}_j|}. \tag{21}$$

where k is the spring constant and l is the equilibrium distance.

A.6. Delfys parameter estimation

Below, we provide the details on how the physical parameters and measurement errors were obtained. All parameters with their errors are shown in Table 4.

Pendulum. The two parameters to approximate are the length of the pendulum L and the damping coefficient ζ. The length L can be trivially measured, and the estimation error is the spacing between marks on the used tape measure, in this case 0.1 cm. Assuming a small initial angle θ_0, the horizontal offset x of the pendulum can be described by:

$$x = A \exp\left(-\frac{\zeta}{2}t\right) \cos(\alpha t - \phi). \tag{22}$$

with A the amplitude, α the frequency, and ϕ the initial phase. The peaks of the above curve then decay as $x_{\text{peak}} = A \exp\left(-\frac{\zeta}{2}t\right)$. The damping coefficient ζ can therefore be obtained by fitting linear regression to the function $\ln(x_{\text{peak}})$. This approximation is done on the setting with the longest length of the string, due to the lowest initial angle θ_0. The error corresponds to the standard deviation over the 5 videos in this setting.

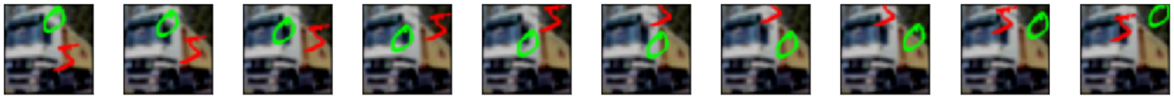


Figure 8. **Dataset baseline**. It shows the evolution of the spring dynamical system of two MNIST digits over a static CIFAR10 background. Figure adapted from [20].

Scenario	**Setting 1**	**Setting 2**	**Setting 3**
Pendulum	$L = 45 \pm 0.1$ [cm]	$L = 90 \pm 0.1$ [cm]	$L = 150 \pm 0.1$ [cm]
	$\zeta = -0.059 \pm 0.008$	$\zeta = -0.059 \pm 0.008$	$\zeta = -0.059 \pm 0.008$
Torricelli	$k = 0.0095 \pm 10^{-4}$ $\left[\frac{\sqrt{m}}{s^2}\right]$	$k = 0.0128 \pm 10^{-4}$ $\left[\frac{\sqrt{m}}{s^2}\right]$	$k = 0.0162 \pm 2 \cdot 10^{-4}$ $\left[\frac{\sqrt{m}}{s^2}\right]$
Sliding block	$\alpha = 20 \pm 1$ [deg.]	$\alpha = 25 \pm 1$ [deg.]	$\alpha = 30 \pm 1$ [deg.]
	$\mu = 0.21 \pm 0.04$	$\mu = 0.21 \pm 0.04$	$\mu = 0.21 \pm 0.04$
LED	$\gamma = 2.3$	$\gamma = 0.92$	$\gamma = 0.46$
Free fall	$r_0 = 3.35 \pm 0.1$ [cm]	$r_0 = 6 \pm 0.1$ [cm]	$r_0 = 10 \pm 0.1$ [cm]
	$h_0 = 20 \pm 0.1$ [cm]	$h_0 = 20 \pm 0.1$ [cm]	$h_0 = 20 \pm 0.1$ [cm]
	$f = 1451 \pm 118$ $\left[\frac{\text{pixels}}{\text{m}}\right]$	$f = 1451 \pm 118$ $\left[\frac{\text{pixels}}{\text{m}}\right]$	$f = 1451 \pm 118$ $\left[\frac{\text{pixels}}{\text{m}}\right]$

Table 4. Physical parameters for each experiment and their measurement errors. The first parameter per scenario is changed between settings.

Torricelli. The only parameter to estimate in this scenario is k relater to the water flow rate. Given the initial height of the water h_0, the final height of the water h_t, and the length of the video t, the parameter k can be computed as:

$$k = 2\frac{\sqrt{h_0} - \sqrt{h_t}}{t}. \tag{23}$$

The video clips were cut such that initial and final heights were correspondingly always $h_0 = 7$cm and $h_t = 1$cm. The error estimate is obtained by computing the standard deviation of the parameter k over the five videos for each setting.

Sliding block. The two parameters to be measured are the inclination angle α and the friction coefficient μ. The inclination angle was set by varying the height of the top of the ramp. Investigating the recorded videos with a protractor showed that the angle was correct to within one degree. Using the sliding block equation as shown in the paper, the friction coefficient can be computed as:

$$\mu = \tan(\alpha) - \frac{2s}{gt^2}. \tag{24}$$

with $s = 72.6$cm the total travel distance of the block and t the duration of the video. Since the friction coefficient should be constant across all settings, the estimate and the error are computed over all 15 videos for this experiment. The error corresponds to the standard deviation.

LED. The only relevant parameter is the decay γ. This decay is controlled automatically, and thus, the value is exact.

Free fall. The radius r_0 is measured with a tape measure with an error of 0.1cm. The initial distance from the camera $h_0 = 20$cm was measured likewise. The focal length if calculated as $f = \frac{r(0)}{r_0} h_0$ and the error is the standard deviation of the focal length over 15 recordings for this experiment. It should be noted that the units of the focal length are $\left[\frac{\text{pixels}}{\text{m}}\right]$ as the focal length makes the conversion between metric and pixel spaces. Also for this reason, the focal length cannot be taken from the producer's spec sheet for the used camera.

A.7. Delfys75 real-video training details

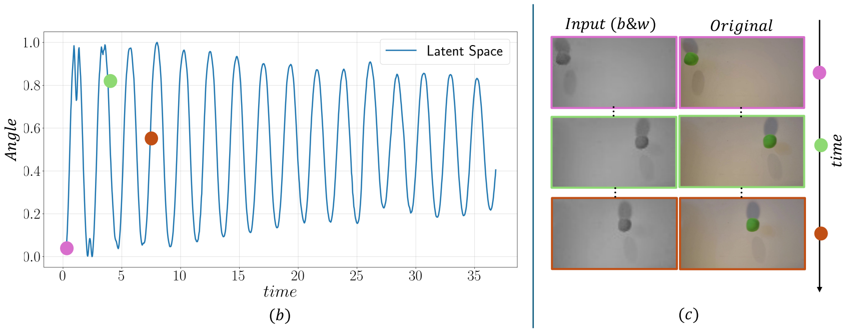
For all videos, we adhered to the specifications outlined in the methods section. The model was trained over 500 epochs with an initial learning rate $lr = 1e^{-2}$. Since each experimental group contained a different number of frames due to varying dynamics, the batch size and the number of input frames per sample were adjusted accordingly. The delta time (dt), defined as the time interval between frames, is determined by the camera's recording speed in frames per second ($fps = 60$). While the minimum possible $dt = \frac{1}{fps}$ was an option, it was not suitable for all experiments. When frame-to-frame differences were negligible ($x_i \approx x_{i+1}$), dt was increased to ensure meaningful variations between frames for prediction.

A.8. Ablation study dt

For this experiment, we examine how changing the Euler step dt affects the training algorithm. Note that, due to the properties of Euler's method, high dt values yield poor approximations, while very small dt values weaken the gradient on the learnable parameter. The result are shown on table 6.

Equation	Parameter	Expected (m)	Estimated (m)
$\theta^{(2)} = -\zeta\theta^{(1)} - \frac{g}{L}\sin(\theta)$	String length (L)	1.20	1.22

(a)



(b)

(c)

Figure 10. Real-world pendulum recording parameter estimation. **(a)** The angle θ is the latent variable. Damping factor ζ and the string length L are learned. **(b)** Extracted dynamics by the model. **(c)** Gray scale input and the original frame from the dataset, related to time in plot (b) using the coloured dots. Our model can estimate the parameter L with only a 0.02 m error.

Experiment	Batch Size	Frames per Sample	$dt\ (s)$
Pendulum	64	20	$\frac{1}{10}$
Torricelli	64	20	$\frac{1}{10}$
Sliding Block	32	10	$\frac{1}{30}$
LED	32	20	$\frac{1}{60}$
Free Fall Scale	all	4	$\frac{1}{30}$

Table 5. Hyperparameters used for training on Delfys75 experiments.

Dataset (Estimated parameter)	dt	Value
Intensity	0.2	0.08
(Damping factor)	0.4	0.078
γ	0.8	0.077
Torricelli	0.1	0.0089
(Flow rate)	0.2	0.0089
$k[\frac{\sqrt{m}}{s}]$	0.4	0.011
Pendulum	0.1	0.45
(String length)	0.2	0.51
$L[m]$	0.4	0.48

Table 6. Estimated parameter for various dt. Results indicate that it has a small impact on predictions.

A.9. Real video latent space visualization

This section presents two different systems: 1. An LED light video recording with the constant brightness change over time 11, similar to the intensity problems previously studied. 2. A pendulum 10 recording which validates the model in a realistic version of the synthetic dataset. For training these models, no masks were needed compared to baselines [17, 20].

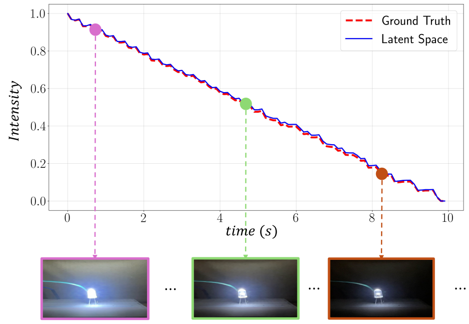


Figure 11. Estimating dynamics from a real-world linear intensity change video. At the top are the estimated dynamics from the model (Latent Space), which are compared with a manually extracted ground truth; the plot shows three dots where example frames of the video inputs are shown in the bottom. The plot shows the latent space can accurately estimate the intensity, capturing the global behaviour over time and following the expected dynamics.

For an LED recording with constant change of brightness over time, 11 shows that the model performs accurately (model predictions in blue) when compared with the ground truth intensity values manually extracted for each frame (red). In Figure 10, we present a more realistic use case where we do not have access to the ground truth or precise manual annotations for each frame. Fortunately, the length of the string L is known to be 120 cm (string length is not visible in the video). The task in this experiment is to learn the value of L. Quantitatively, in Figure 10a, the model can reliably estimate the value of the length parameter L. Besides, in the latent space Figure 10b, we can see that the model can accurately predict the natural damped oscillations.