VIVALDy: AI-Driven Low-Order Modeling of Vortex-Induced Vibrations via β-Variational Autoencoders, Transformers, and Adversarial Training

Niccolò Tonioni, Lionel Agostini, Franck Kerhervé, Laurent Cordier, Ricardo Vinuesa

VIVALDy: A Hybrid Generative Reduced-Order Model for Turbulent Flows, Applied to Vortex-Induced Vibrations

A PREPRINT

Niccolò Tonioni[*1], **Lionel Agostini**[1], **Franck Kerhervé**[1], **Laurent Cordier**[1], **and Ricardo Vinuesa**[2,3]

[1]Pprime Institute, CNRS, Université de Poitiers, ISAE-ENSMA,Chasseneuil-du-Poitou, 86360, France
[2]Department of Aerospace Engineering,University of Michigan, Ann Arbor, MI 48109, United States
[3]FLOW, Engineering Mechanics, KTH Royal Institute of Technology, Stockholm, SE-100 44, Sweden

November 12, 2025

ABSTRACT

Developing reduced-order models applicable to fluid-dynamics problems involving complex geometries and different flow conditions remains a critical challenge for turbulent flows. This study introduces VIVALDy, a novel machine-learning framework that employs a hybrid β-Variational Autoencoder-Generative Adversarial Network (β-VAE-GAN) architecture with masked convolutions to extract dominant flow features into a compact latent space while preserving fidelity at solid-fluid interfaces. A bidirectional transformer then models the temporal evolution of these features, learning to predict flow trajectories from minimal sensor inputs. This two-stage approach enables the transformer to map sensor measurements to dominant flow variables identified by the autoencoder, advancing reduced-order modeling capabilities for real-time flow prediction. The effectiveness of the framework is demonstrated through application to a problem relevant to vortex-induced vibration (VIV) energy harvesting systems, reconstructing the turbulent flow around a one-degree-of-freedom moving cylinder. Validated against experimental data spanning fluid-structure interaction regimes of interest, VIVALDy accurately predicts different flow states using only the cylinder displacement. The framework demonstrates adequate performance in both reconstruction accuracy and statistical fidelity across diverse operating conditions, enabling efficient prediction of the turbulent flow phenomena governing vortex-induced vibration.

Keywords Reduced-Order Models · Turbulent Flows · Vortex-Induced Vibrations · β-Variational Autoencoders · Generative Adversarial Networks · Transformers

1 Introduction

The increasing concerns about the global climate crisis and the United Nations Sustainable Development Goal 7 (*ensuring access to affordable and clean energy*) have stimulated significant interest among public institutions in renewable energy-harvesting technologies [1]. While the wind energy sector has mature technologies, wave and tidal energy remain in the proof-of-concept stage, limiting exploitation of Earth's vast water current resources [2]. These technologies face unique challenges in marine and fluvial ecosystems: devices must maintain efficiency under varying flow conditions, and be unobtrusive with low maintenance costs due to biofouling. Vortex-induced vibration (VIV) systems offer a promising solution due to their simple structure and operation mechanism [3, 4].

VIV is a form of self-excited fluid-structure interaction (FSI) characteristic of bluff bodies, such as cylinders, immersed in a fluid flow [5]. VIV energy harvesting systems exploit these structural vibrations rather than suppress them, converting mechanical motion into usable energy. However, to maximize energy extraction from these devices under

[*]niccolotonioni.github.io, niccolo.tonioni@univ-poitiers.fr

all operating conditions, real-time control strategies and rapid design optimization are essential. Therefore, there is a need for the development of Reduced-Order Models (ROMs) that can overcome the prohibitive computational costs associated with predicting the turbulent flow fields that govern VIV phenomena.

ROMs are techniques designed to create computationally efficient surrogate models by constructing low-dimensional approximations preserving the most informative features of the original system [6]. Historically, research on ROMs has focused on linear methods with two main classes emerging: projection-based ROM (PB-ROM) and inference-based ROM (IB-ROM). PB-ROMs are inherently intrusive as they require explicit basis expansions and projections of the full-order model operators [7, 8]. In contrast, IB-ROMs offer a non-intrusive alternative. They infer surrogate operators directly from the data based on prior knowledge of governing equation structure, without requiring access to full-order operators [9, 10]. However, they may require regularization techniques as they are based on solving least-squares problems, which can be ill-conditioned and sensitive to noise [11]. Furthermore, constructing full-order operators remains computationally expensive, necessitating data projections with associated truncation errors. Non-linear IB-ROMs have been developed to reduce truncation errors and improve accuracy [12, 13]. However, approximating both linear and quadratic operators of the Navier-Stokes equations may incur even higher computational costs. Another limitation of these models is their implementation, which is traditionally in a non-parametric form. Consequently, they demonstrate an inability to generalize to variations in flow conditions, such as changes in incoming flow velocity a particular area of interest in the present study. Although some studies have explored how to extend these classical models to parameterized dynamic problems [14, 15], this area has seen only limited development.

To address these limitations, researchers have increasingly explored machine-learning techniques for ROM development. The objective is developing non-linear approaches that improve accuracy, efficiency, and generalizability across flow configurations [16, 17]. Autoencoders (AEs) have gained particular attention due to their success in areas such as image compression. An AE, first introduced by Hinton and Salakhutdinov [18], is a neural network architecture comprising two primary (non-linear) mappings: an encoder and a decoder. The encoder compresses high-dimensional input data into lower-dimensional latent representations, also termed code. Subsequently, the decoder reconstructs the input data from this latent space, returning to the original high-dimensional space. The method is purely data-driven, with the network weights optimized by minimizing an objective function. Consequently, Autoencoders offer a non-intrusive, equation-free framework for constructing models that generalize across diverse flow parameters, provided these parameters are represented in the training dataset.

Agostini explored the application of AE for modeling a two-dimensional laminar cylinder wake flow [19]. In the study, the author compares the results obtained using Proper Orthogonal Decomposition (POD), a classical PB-ROM approach, with those obtained using AE for a latent space of dimension three. The results indicate that the AE surpasses POD in terms of reconstruction quality. However, a limitation of AE is the lack of orthogonality of the low-dimensional representations. Inspired by the Cluster-ROM algorithm [20], Agostini also constructed a probabilistic dynamical model in the AE latent space using spectral clustering and Markov chains. This statistical model allowed the author to quantify the likelihood of specific clusters occurring under given flow states. Since each cluster corresponds to distinct dynamics within the latent space and is associated to specific flow states, this approach provided physical interpretability of the underlying flow behavior.

Eivazi *et al.* [21] extended this line of search to Variational Autoencoders (VAEs), which generalize AEs through stochastic mappings for low-dimensional representation of flow dynamics. Specifically, the authors employed a β-VAE [22], a modified VAE architecture that introduces a hyperparameter β to balance reconstruction fidelity against latent space regularization toward a standard normal distribution. This variation enables the user to control the model's training to prioritize *disentanglement* in the latent space, i.e., to separate independent factors of variation in the data. Eivazi *et al.* [21] applied a β-VAE to flows in a simplified urban environment, and compared the results against AE and POD. Their results demonstrate that the β-VAE not only achieves better reconstruction of the flow fields for a given compression ratio, but it also achieves more orthogonal latent variables compared to AE.

In a series of subsequent papers, Wang *et al.* [23] and Solera-Rico *et al.* [24] combined β-VAEs with other neural architectures to model the temporal dynamics in the latent space and develop ML-driven ROMs for predicting turbulent flows. Specifically, they compared the performance of Long Short-Term Memory (LSTM) [25], a recurrent architecture introduced for learning long-range dependencies in sequence data, with Transformers [26], which, in contrast, exploit self-attention mechanisms to model dependencies without relying on recurrence, offering advantages in parallelization. Both studies concluded that Transformers have lower errors and the best long-term reconstruction performance.

Inspired by previous works, this paper introduces VIVALDy, a novel ML-based ROM framework (see Figure 1 for its overall structure). In contrast to existing methodologies, VIVALDy can reconstruct the flow state across varying flow conditions from minimal sensor measurments. When applied against a newly acquired experimental dataset encompassing a range of fluid-structure interaction (FSI) regimes directly relevant to practical VIV energy-harvesting applications, VIVALDy successfully reconstructs the turbulent flow around a one-degree-of-freedom cylinder using only

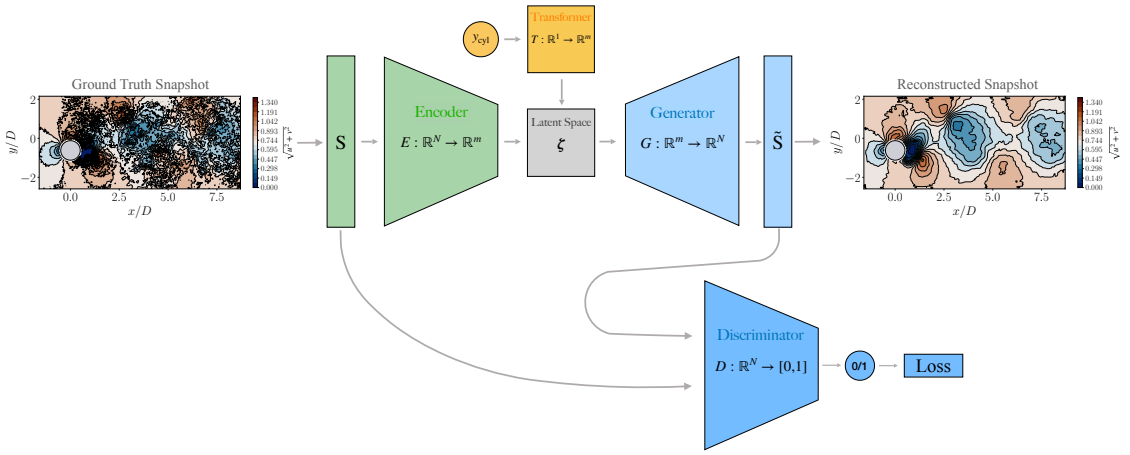


Figure 1: Schematic representation of VIVALDy framework. *Training phase:* The β-variational autoencoder (β-VAE) and discriminator are trained simultaneously in a generative adversarial framework, where the decoder serves as generator and receives evaluative feedback from the discriminator. A transformer model is then trained to predict latent variable evolution using only cylinder displacement y_{cyl} as input. *Inference phase:* Only the transformer and decoder are retained to generate flow field predictions from displacement signals.

the cylinder displacement as input. Moreover, the model's predictions are both computationally and memory-efficient as it operates directly in the latent space.

From a physical point of view, VIVALDy's approach is feasible due to the coupling between the cylinder's motion and the surrounding flow field, resulting in correlated signals. From a computational perspective, this is made possible by combining a novel hybrid generative architecture for latent-feature extraction with a bidirectional transformer designed to learn the non-linear correlations between the cylinder kinematics and the flow dynamics. The hybrid architecture, which is based on β-VAE and Generative Adversarial Networks (GANs) [27], is motivated by recent works in image compression [28, 29, 30]. These works demonstrate that this combination leverages the individual strengths of both models, resulting in improved distribution-preserving properties compared to a standard VAE and a more structured, informative latent space than that offered by a standalone GAN.

The paper is organised as follows: Section 2 details the experimental dataset and its preparation. Section 3 provides a detailed description of the VIVALDy framework. Section 4 presents the principal results, which are discussed in Section 5. Finally, Section 6 summarizes the main contributions and discusses future outlook.

2 Dataset Generation and Preparation

This section details the experimental dataset employed for the training and validation of VIVALDy. The dataset comprises streamwise and crosswise velocity fields (u, v) of an elastically mounted cylinder undergoing vortex-induced vibrations in the cross-flow direction. The data were acquired using time-resolved Particle Image Velocimetry (PIV). The following subsections describe the experimental setup and the preparation of the acquired flow snapshots for model input.

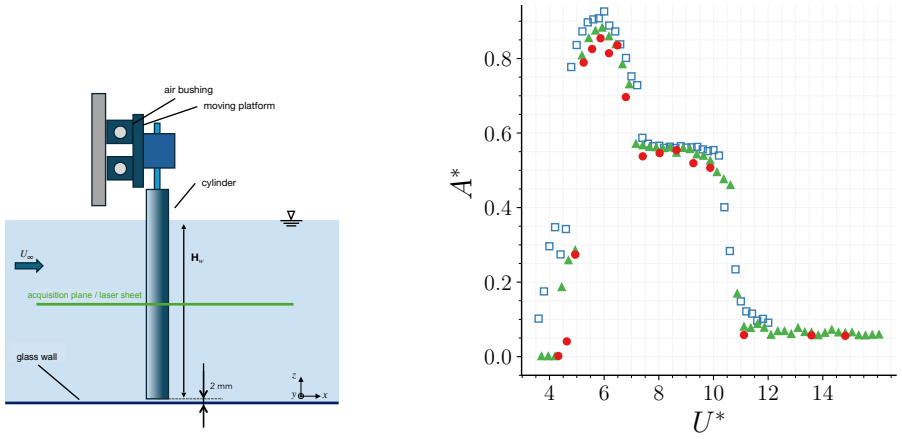
2.1 Experimental Setup and Data Acquisition

The experiments were conducted in the closed-circuit water channel Hydra III at Institut Pprime (Poitiers, France). Figure 2a presents a schematic illustration of the experimental setup. The test section measures $2.1\,\mathrm{m}$ in length, $0.51\,\mathrm{m}$ in width, and $0.51\,\mathrm{m}$ in height. The cylinder, which has a length of $0.4\,\mathrm{m}$ and a circular cross-section of diameter $D = 0.05\,\mathrm{m}$, is mounted on an elastically supported platform with negligible damping, allowing it to oscillate freely in the cross-flow direction. The maximum allowed displacement of the cylinder is $\pm 2D$.

The PIV measurement plane is positioned at mid-height ($H_w/2$) of the immersed portion of the cylinder and covers a region of $4D \times 9D$ in the crosswise (y) and streamwise (x) directions, respectively. The velocity fields were acquired at a sampling frequency $f_s = 10\,\mathrm{Hz}$, which is an order of magnitude larger than the maximum dominant shedding

frequency observed in the data, thereby guaranteeing time-resolved flow structures. The grid spatial resolution is $\Delta x = \Delta y = 0.025D$, ensuring spatially resolved flow measurements.

A total of $N_{\text{oc}} = 17$ operating conditions were measured. Figure 2b depicts these conditions on the characteristic amplitude response plot of the cylinder, defined by the normalized amplitude $A^* = A/D$ as a function of reduced velocity $U^* = U_\infty / f_n D$, with U_∞ the inflow velocity and $f_n = 0.421$ Hz the cylinder natural oscillation frequency. This response is typical of a low-mass ratio VIV system, and the acquired operating conditions cover all the distinct FSI regimes of interest to VIV energy harvesting devices [31, 4, 32]. Further details about these regimes are reported in Section 2.2. A detailed description of the instrumentation and VIV device can be found in Schmider et al. [4].



(a) (b)

Figure 2: Experimental setup and cylinder amplitude response. (a) Side-view of the experimental test section setup for Particle Image Velocimetry (PIV) measurements, showing the laser sheet position. (b) Characteristic amplitude response $A^* = A/D$ of the cylinder as a function of reduced velocity $U^* = U_\infty / f_n D$. The red dots indicate the 17 cases acquired in the present study, while results from Schmider et al. [4] (green triangles) and Soti et al. [32] (blue squares) are included for comparison.

2.2 Preparation of 2D Flow Snapshots for Model Input

The acquired PIV measurements define a dataset $\mathcal{D}$ of two-dimensional (x-y) snapshots including the streamwise and crosswise velocity fields (u-v). A total of $N_t = 1,000$ snapshots were extracted for each of the $N_{\text{oc}} = 17$ operating conditions shown in Figure 2b. The dataset can therefore be written as:

$$\mathcal{D} = \bigcup_{i=1}^{N_{\text{oc}}} \mathcal{D}_i \,, \quad \text{where} \quad \mathcal{D}_i = \left\{ \mathbf{S}_j^{(i)} \right\}_{j=1}^{N_t} \tag{1}$$

with the snapshot $\mathbf{S}_j^{(i)}$ being a structured tensor of shape $N_x \times N_y \times N_c = 416 \times 194 \times 2$. As an extension of this notation, the superscripts Train, Val and Test will be used to denote the data corresponding to the training, validation and test sets defined later in this section.

The raw PIV data often contain spurious missing values and exhibit long tails in their probability density function, issues that can compromise the network training. To mitigate these, a two-step preprocessing procedure was applied to the PIV acquisitions. For each operating condition i, the dataset $\mathcal{D}_i$ is first filtered with Singular Value Decomposition (SVD). Specifically, the singular values were truncated to retain 99% of the relative information content, quantified as the ratio of the sum of the retained singular values to the total sum. Second, to remove extreme outliers the probability density functions of each velocity component c in $\mathcal{D}_i$ were clipped at $\mu_c^{(i)} \pm 3\sigma_c^{(i)}$. Here, $\mu_c^{(i)}$ and $\sigma_c^{(i)}$ are the ensemble mean and standard deviation of the channel c computed as:

$$\mu_c^{(i)} = \frac{1}{N_t N_x N_y} \sum_{j,k,p} S_{j,k,p,c}^{(i)} \quad , \quad \sigma_c^{(i)} = \sqrt{\frac{1}{N_t N_x N_y} \sum_{j,k,p} \left(S_{j,k,p,c}^{(i)} - \mu_c^{(i)} \right)^2} \quad . \tag{2}$$

This clipping threshold retains 99.73% of the original probability density function, assuming a Gaussian distribution.

After this two-step preprocessing the dataset $\mathcal{D}$, containing a total of $17,000$ snapshots, was partitioned into three subsets: a training set $\mathcal{D}^{\text{Train}}$, a validation set $\mathcal{D}^{\text{Val}}$, and a test set $\mathcal{D}^{\text{Test}}$. The training set comprises the first $N_t^{\text{Train}} = 900$ snapshots from the $N_{\text{oc}}^{\text{Train/Val}} = 12$ different operating conditions represented in orange in Figure 3a (10,800 snapshots in total). The subsequent $N_t^{\text{Val}} = 100$ snapshots from these same operating conditions form the validation set (a total of 1200 snapshots), employed for hyperparameter selection, monitoring training progress against overfitting, and implementing early stopping criteria. Finally, the test set consists of all $N_t^{\text{Test}} = 1,000$ snapshots from the $N_{\text{oc}}^{\text{Test}} = 5$ different operating conditions represented in blue in Figure 3a (a total of $5,000$ snapshots), strictly reserved for evaluating the performance of the final trained models (results presented in Section 4).

This test set was designed to be entirely separate from the training and validation sets (orange cases) to rigorously evaluate the model's generalization capability to unseen flow dynamics. Additionally, the test cases were specifically selected to encompass the characteristic fluid-structure interaction regimes present in the dataset:

- Initial Branch: This regime occurs at lower reduced velocities, where the vortex shedding frequency is below the structure's natural frequency. The cylinder remains nearly stationary, despite sensing fluid forces.
- Upper Branch: As the reduced velocity increases, the vortex shedding frequency begins to match the cylinder's natural frequency, thereby inducing a resonance phenomenon that results in the maximum recorded cylinder oscillation amplitude.
- Transition Branch: Further increasing the reduced velocity leads to a shedding frequency slightly higher than the structure's natural frequency. An intermittent regime characterized by switching behavior between the Upper and Lower Branches is therefore observed.
- Lower Branch: As reduced velocity (and thus shedding frequency) increases, the cylinder's oscillations synchronize with the shedding frequency (lock-in), following the flow instability.
- Asynchrony Region: In this region, the shedding frequency significantly exceeds the natural frequency, leading to a weak coupling of the fluid and structural dynamics. The cylinder displays minimal oscillations.

Figure 3b helps visualize how the flow patterns differ across these distinct branches. It is important to note that the Transition regime is exclusively present in the test set, a higher generalization challenge is thus anticipated for this specific regime. Nevertheless, given that the primary flow mechanisms in this branch represent a combination of those found in the Upper and Lower Branches, the model is still expected to demonstrate generalization capabilities.

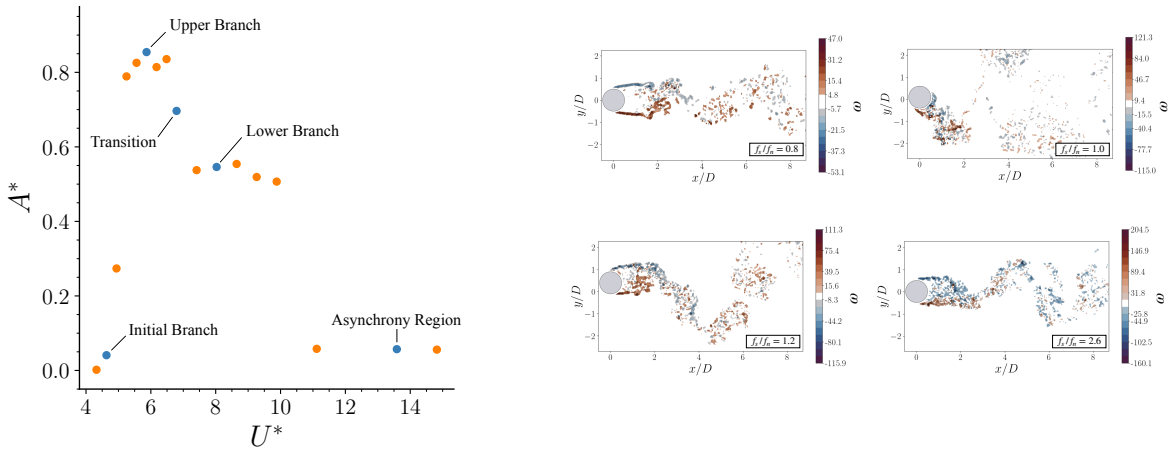
3 VIVALDy: A Machine-Learning Framework for Data-Driven Flow Modeling

This section details the proposed VIVALDy machine-learning framework, describing its components and architectures.

3.1 Masked Convolutions for Solid-Fluid Interface Fidelity

A primary challenge in applying convolutional neural networks (CNNs) to fluid domains is handling the presence of structural elements within the convolution field of view, as missing values cannot be backpropagated during training [33]. Encoding and reconstructing these interface regions is essential as this is where flow instabilities are generated and fluid-structure interaction occurs [34, 35]. Excluding such regions would result in loss of relevant flow information during encoding, leading to unsatisfactory reconstructions.

To illustrate, considering the case under study, the presence of the oscillating cylinder must be addressed. A naive approach would be masking the cylinder's grid points with zeros. However, this approach would create ambiguity for the network: a zero value could represent either an actual physical zero (e.g., zero velocity at a stagnation point) or a missing data point within the solid body. Since the cylinder's position changes dynamically across snapshots, this ambiguity might force the network to learn non-physical interpretations. Additionally, since the PIV data exclude the boundary layer, zero masking introduces sharp discontinuities that cause high gradients during backpropagation, leading to unstable training or incorrect inferences. Therefore, this work employs masked CNNs [36], originally developed for sparse depth estimation, which utilize specialized convolutions with binary masks to distinguish between fluid and solid



(a) (b)

Figure 3: Dataset partitioning strategy and illustrative VIV flow patterns. (a) Amplitude response plot ($A^* = A/D$) versus reduced velocity ($U^* = U_\infty/f_n D$), delineating cases used for training/validation (orange dots) and the test set cases (blue dots). (b) Vorticity ($\omega = \partial v/\partial x - \partial u/\partial y$) visualizations illustrating the diverse fluid-structure interaction regimes within the test set. The normalized vortex shedding frequency (f_s/f_n) is also reported for each case. From top-left: Initial, Upper, Lower, and Asynchrony Branches.

regions. This approach enables the network to explicitly account for the cylinder's presence and focus calculations solely on the fluid domain.

Formally, let $\mathbf{o} = \{o_{i,j}\}$ be the output of the previous layer and $\mathbf{m} = \{m_{i,j}\}$ the binary mask ($m_{i,j} = 1$ if the point of coordinate (i, j) belongs to the fluid and $m_{i,j} = 0$ if the point belongs to the solid). The masked convolution operation is defined as:

$$f_{p,q}(\mathbf{o}, \mathbf{m}) \;=\; \frac{\sum_{i,j=-k}^{k} m_{p+i,q+j}\, o_{p+i,q+j}\, w_{ij}}{\sum_{i,j=-k}^{k} m_{p+i,q+j} + \varepsilon} \;+\; b, \tag{3}$$

where w_{ij} are the convolution weights, k is the kernel size, b is a bias term, and ε is a small constant used to avoid division by zero. When all points under the convolution kernel belong to the fluid (i.e. all $m_{p+i,q+j}$ are non-zero), this equation reduces to a standard convolution. The mask $\mathbf{m}$ itself is updated through subsequent layers by max pooling, which allows the network to maintain a consistent notion of valid (fluid) and invalid (solid) regions as the data passes through the network:

$$f^{\mathbf{m}}_{p,q}(\mathbf{m}) \;=\; \max_{i,j=-k,\ldots,k} m_{p+i,q+j}, \tag{4}$$

where k is also the kernel size for the max pooling operation.

3.2 Latent-Feature Extraction using a β-VAE-GAN Architecture

The core architecture of VIVALDy is now presented, comprising a latent-feature extractor designed to extract informative low-dimensional latent features $\boldsymbol{\zeta}$ from high-dimensional flow snapshots $\mathbf{S}$. This architecture incorporates masked convolutions, which effectively manage the solid-fluid interface. The vector $\boldsymbol{\zeta}$ is defined to have a size of m, which is a predefined parameter. In this study, the model compresses each $416 \times 194 \times 2$ flow snapshot, defined in Section 2.2, into a latent space of just $m = 3$ dimensions, representing an extreme compression ratio exceeding $50,000 : 1$.

This extreme compression is achieved while preserving flow statistical properties using a hybrid generative architecture. The proposed design merges β-Variational Autoencoders [22] with Generative Adversarial Networks [27], motivated by recent works in learned image compression [28, 29, 30]. These works showed how adding adversarial GAN loss to

autoencoder-based compression models improves input data distribution preservation in reconstructions. These improvements are obtained through simultaneous optimization of three objectives: reconstruction fidelity, disentanglement of latent variables, and statistical consistency between reconstructed and reference distributions, as will be explained. The following briefly summarizes the foundational components of the proposed architecture.

VAEs are a type of autoencoder that learns *probabilistic* latent-space representations of data. The β-VAE modification introduces a scaling factor β to the Kullback–Leibler (KL) divergence term in the objective function. Assuming the prior distribution $p_{\boldsymbol{\zeta}}$ follows a standard multivariate normal distribution, the encoder E approximates the posterior $p_{\boldsymbol{\zeta}|\mathbf{S}}$ via a probability $q_{\mathbf{W}_E}(\boldsymbol{\zeta}|\mathbf{S})$, parametrized by the encoder weights $\mathbf{W}_E$. The decoder G then acts as a generative model, approximating $p_{\mathbf{S}|\boldsymbol{\zeta}}$ through a distribution $p_{\mathbf{W}_G}(\mathbf{S}|\boldsymbol{\zeta})$ parametrized by its weights $\mathbf{W}_G$. The optimization process minimizes a modified evidence lower bound (ELBO), augmented by β:

$$\mathcal{L}_{EG}^{\beta-\mathrm{VAE}} = \mathbb{E}_{\mathbf{S}\sim p_{\mathbf{S}}}\big[d(\mathbf{S},\tilde{\mathbf{S}})\big] + \beta\, D_{\mathrm{KL}}\big(q_{\mathbf{W}_E}(\boldsymbol{\zeta}|\mathbf{S})\,\|\,p_{\boldsymbol{\zeta}}\big), \tag{5}$$

where where $\mathbb{E}[\cdot]$ denotes expectation, $d(\cdot,\cdot)$ is a distortion metric, e.g. the mean square error, and $D_{\mathrm{KL}}(\cdot)$ is the Kullback–Leibler divergence.

The β parameter provides a mechanism to balance the trade-off between reconstruction fidelity $d(\cdot,\cdot)$, and disentanglement in the latent space $D_{\mathrm{KL}}(\cdot)$. Disentanglement refers to the ability to separate independent factors of variation in the data, allowing each latent dimension to capture distinct underlying features. In the context of physical systems, this property enables the identification of independent physical parameters that generated the observed data [37]. Higher values of β promote greater disentanglement; however, this process may potentially result in a decrease in reconstruction fidelity.

Generative Adversarial Networks are probabilistic models composed of two networks, a generator G and a discriminator D, trained in opposition to one another. The generator $G(\boldsymbol{\zeta})$ maps latent vectors drawn from a standard Gaussian $\boldsymbol{\zeta}\sim p_{\boldsymbol{\zeta}}$ to reconstructions $\tilde{\mathbf{S}}$, while the discriminator D attempts to distinguish between input samples $\mathbf{S}\sim p_{\mathbf{S}}$ and generated samples $\tilde{\mathbf{S}}\sim p_{\tilde{\mathbf{S}}}$. This two-player min-max game is formalized using the following non-saturating loss:

$$\mathcal{L}_G^{\mathrm{GAN}} = \mathbb{E}_{\boldsymbol{\zeta}_j\sim p_{\boldsymbol{\zeta}}}\Big[-\log\big(D\big(G(\boldsymbol{\zeta})\big)\big)\Big], \tag{6}$$

$$\mathcal{L}_D^{\mathrm{GAN}} = \mathbb{E}_{\mathbf{S}\sim p_{\mathbf{S}}}\Big[-\log\big(D(\mathbf{S})\big) - \log\big(1-D\big(\tilde{\mathbf{S}}\big)\big)\Big]. \tag{7}$$

The generator's loss, $\mathcal{L}_G^{\mathrm{GAN}}$, represents the negative log probability that the discriminator assigns to the generated data $\tilde{\mathbf{S}} = G(\boldsymbol{\zeta})$ being from the real data distribution. Conversely, the discriminator's loss, $\mathcal{L}_D^{\mathrm{GAN}}$, minimizes two terms: the negative log probability that it assigns to real data $\mathbf{S}$ being real (first term), and the negative log probability that it assigns to generated data being fake $\tilde{\mathbf{S}}$ (second term). This adversarial loss formulation, where the generator tries to minimize $\mathcal{L}_G^{\mathrm{GAN}}$ and the discriminator tries to minimize $\mathcal{L}_D^{\mathrm{GAN}}$, drives the generator to produce outputs that are statistically consistent with the real data distribution, $p_{\tilde{\mathbf{S}}}\approx p_{\mathbf{S}}$.

The combined β-VAE-GAN framework (illustrated in Figure 4) consists of three components: an encoder E, a decoder/generator G, and a discriminator D. The *masked convolutions*, introduced in the previous subsection are employed in the encoder and discriminator, to explicitly account for the cylinder presence in the flow domain. In contrast, the decoder does not require masking, as the masked convolutions in the encoder ensure that only the fluid region grid points contribute to its inputs (the latent variables representing the flow region). The encoder and generator are jointly optimized, while being trained adversarially against the discriminator. The training objective integrates terms from both the β-VAE and the GAN formulations into the following adversarial loss:

$$\mathcal{L}_{EG}^{\mathrm{Hybrid}} = \underbrace{\mathbb{E}_{\mathbf{S}\sim p_{\mathbf{S}}}\big[d(\mathbf{S},\tilde{\mathbf{S}})\big] + \beta\, D_{\mathrm{KL}}\big(q_{\mathbf{W}_E}(\boldsymbol{\zeta}|\mathbf{S})\,\|\,p_{\boldsymbol{\zeta}}\big)}_{\beta\text{-VAE terms}} - \alpha\,\underbrace{\mathbb{E}_{\mathbf{S}\sim p_{\mathbf{S}}}\big[\log\big(D(\tilde{\mathbf{S}})\big)\big]}_{\text{GAN term}}, \tag{8}$$

$$\mathcal{L}_D^{\mathrm{Hybrid}} = \underbrace{\mathbb{E}_{\mathbf{S}\sim p_{\mathbf{S}}}\big[-\log\big(D(\mathbf{S})\big) - \log\big(1-D\big(\tilde{\mathbf{S}}\big)\big)\big]}_{\text{GAN term}}. \tag{9}$$

Here, β and α are user-defined hyperparameters that control the trade-off between reconstruction accuracy $d(\cdot,\cdot)$, latent disentanglement $D_{\mathrm{KL}}(\cdot)$, and statistical fidelity $\log D(\cdot)$. During training, two separate optimizers update the networks parameters: one for the combined EG loss (Equation 8) and another one for the discriminator (D) loss (Equation 9). Training details are provided in Section 3.4.

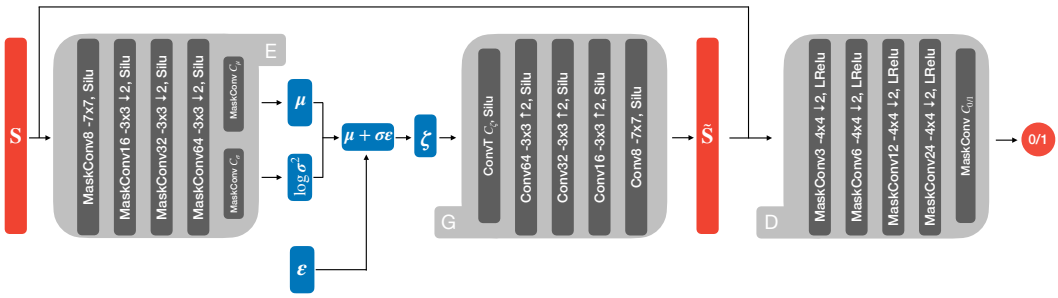


Figure 4: β-VAE-GAN architecture. The encoder (E), generator (G), and discriminator (D) use (masked) convolutional layers. Each (Mask)Conv$C$ layer applies $k \times k$ filters with C output channels, where k is the filter kernel size. Downward arrows ($\downarrow 2$) denote strided down-sampling, while upward arrows ($\uparrow 2$) denote Lanczos up-sampling. Silu (*sigmoid linear unit*) and LRelu (*leaky ReLU*) are used as activation functions. The layers labelled `Conv` C_μ and `Conv` C_σ project their inputs to the mean and variance of the latent variables, respectively, whereas `ConvT` C_ζ projects the sampled latent variable $\boldsymbol{\zeta}$ back to the encoder's final layer dimensionality.

3.3 Latent-Space Dynamics from Cylinder Kinematics using a Bidirectional Transformer

The β-VAE-GAN architecture provides spatial compression for each snapshot. The physical system dynamics are consequently represented as an evolving trajectory within the low-order latent space. To reconstruct the full spatio-temporal dynamics from cylinder displacement alone, a separate architecture that maps this displacement directly to the target trajectory in the latent space is introduced. When coupled with the β-VAE-GAN decoder, this architecture enables reconstruction of the original high-dimensional flow fields.

The model's objective is to learn a mapping, T, that predicts the flow's trajectory in the low-dimensional latent space from the cylinder's displacement.

Given an input sequence of cylinder displacements over a time window of length H, denoted as $y_{cyl} = [y_{t-H+1}, \ldots, y_t]$, the model outputs the corresponding trajectory in the latent space, $\zeta = [\zeta_{t-H+1}, \ldots, \zeta_t]$. Each vector ζ_t in this sequence (for example, $\zeta_t = [\zeta_{t,1}, \zeta_{t,2}, \zeta_{t,3}]$) represents the compressed state of the flow at that moment. The complete function is defined as:

$$\boldsymbol{\zeta} = T(\mathbf{y}_{\text{cyl}}). \tag{10}$$

Physically, this mapping T can be interpreted as the learned coupling between the structural motion of the cylinder and the dynamics of the surrounding flow field.

The problem is solved using an encoder-only bidirectional transformer (Figure 5), drawing inspiration from Bidirectional Encoder Representations from Transformers (BERT) [38]. The architecture comprises three primary components: a time-embedding (TE), a bidirectional transformer block (BTB), and a linear projection (LP). The TE component consists of a two-layer CNN that encodes the input displacement sequence to a higher-dimensional representation of dimension n_e. Positional encoding is then added to provide the sequence order information, and the resulting tensor is passed to the BTB. The BTB consists of six transformer encoder layers, each containing six-head self-attention mechanisms combined with feedforward networks. The self-attention mechanism operates bidirectionally without masking, allowing each time step to simultaneously consider information from all other time steps in the sequence, following the original transformer encoder formulation by Vaswani *et al* [26]. Finally, the output passes through a linear projection layer, consisting of a one-layer perceptron, which maps the BTB output to the desired latent space trajectory dimension. Two hyperparameters control the model configuration: the TE output dimension n_e and a parameter n_H defining the maximum context length. The latter specifies the maximum number of time steps the attention mechanism can process simultaneously. While the model can handle sequences of arbitrary length H through windowing strategies, attention computations are constrained to this maximum context length. Further training and hyperparameter details are reported in Section 3.4.

The design choice of using bidirectional over unidirectional attention is better understood when considering the simplified scheme presented in Figure 6. The objective is to learn the mapping between output and input time series. When mapping a general time instant t_i (for example t_4 in the figure), a unidirectional attention mechanism constrains

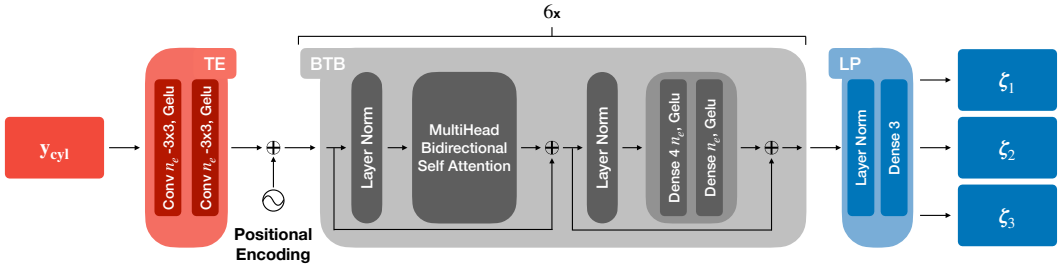


Figure 5: Bidirectional transformer architecture. The time embedding (TE) layer consists of two convolutional layers with 3×3 filters and Gelu (Gaussian error linear unit) activation functions. The number of filter channels is set to n_e, which determines the dimensionality of the encoded representation fed into the bidirectional transformer block (BTB). The latter uses a multihead, bidirectional self-attention mechanism, following the original transformer formulation [26], combined with a shallow dense feedforward network. This transformer block is repeated six times. Finally, a dense latent projection (LP) layer maps the transformer's output to the latent trajectory space.

the output t_i to depend only on input elements prior to its time step (t_j with $j < i$). This constraint becomes problematic when significant temporal lags or leads exist between the input time series (cylinder kinematics) and the output time series (latent flow dynamics), as it prevents the model from accessing the full-context information. In contrast, a bidirectional attention mechanism operates without such masking, allowing the model to utilize information from the complete input time series. Additionally, if the underlying problem exhibits inherent unidirectional dependencies, the bidirectional mechanism can converge to unidirectional behavior during training. By utilizing the full temporal context of the cylinder's displacement signal, this approach enhances the predictive accuracy of latent-space trajectories.

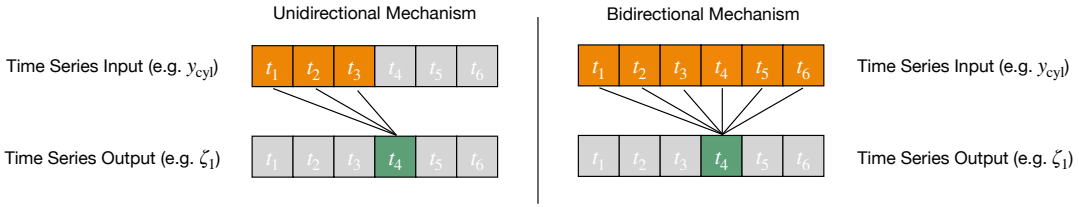


Figure 6: Comparison of unidirectional and bidirectional attention mechanisms for time series. In a unidirectional mechanism (left), the output at time t_4 can only attend to inputs before t_4 (e.g., t_j where $j < 4$). This masking restricts the model's ability to capture lead-lag relationships. Conversely, a bidirectional mechanism (right) allows the output at t_i to attend to all input timesteps.

3.4 Models Training Details

All hyperparameters used during the training of the β-VAE-GAN and bidirectional transformer are detailed in Table 1.

The β-VAE-GAN architecture was trained on $\mathcal{D}^{\text{Train}}$ (Section 2.2) using a mini-batch ADAM optimizer with random shuffling. The model was trained for $2,500$ epochs on a single NVIDIA V100 32GB GPU on the Jean-Zay CNRS supercomputer. During training, the learning rate was adjusted using a warmup cosine decay schedule, and the β hyperparameter was cyclically annealed [39]. An early stopping strategy was adopted to prevent overfitting by monitoring the loss on $\mathcal{D}^{\text{Val}}$. The discriminator weights were updated every 5 epochs and frozen during the remaining epochs to stabilize the GAN training dynamic and prevent the discriminator from dominating the generator. The hyperparameter α was kept constant throughout training. An ablation study on this parameter was conducted, training models with $\alpha = \{0.2, 0.02, 0.002\}$ and comparing against an equally trained β-VAE, which served as a benchmark, to assess its impact on performance. The results of this α sensitivity study are reported in Section 4.1.

The trained β-VAE-GAN encoder (specifically, the $\alpha = 0.2$ model which showed the best performance in the ablation study) was then used to encode $\mathcal{D}^{\text{Train}}$ into the learned low-dimensional space. The obtained target trajectories $\boldsymbol{\zeta}^{\text{gt}}$ were divided into target sequences of size $H = 64$. The same partitioning was applied to the corresponding cylinder displacement acquired during the experiment. The obtained input/output pairs were used to define the training dataset

Table 1: Optimizer hyperparameters used for training of the β-VAE-GAN (Figure 4) and bidirectional-transformer (Figure 5).

Optimizer Encoder/Generator			
Batch Size (N_b)	32	Warmup Steps (ws)	100
Initial learning rate (η_{start})	2×10^{-2}	Initial β (β_{start})	0.0
Final learning rate (η_{end})	2×10^{-5}	Final β (β_{end})	1×10^{-4}
Number of cycles (N_{cycles})	2	Cycle ratio (R)	0.5
Optimizer Discriminator			
Batch Size (N_b)	32	Warmup Steps (ws)	100
Initial learning rate (η_{start})	2×10^{-2}	Final learning rate (η_{end})	2×10^{-5}
Optimizer Transformer			
Batch Size (N_b)	64	Warmup Steps (ws)	100
Initial learning rate (η_{start})	3×10^{-4}	Final learning rate (η_{end})	3×10^{-8}
Dropout rate (dr)	2×10^{-1}		

for the bidirectional transformer (Figure 5). The model context length was set to $n_H = 64$, while the temporal encoding dimension $n_e = 36$. The model was optimized using a mini-batch ADAM optimizer with a warmup cosine decay schedule for the learning rate. An early stopping strategy was adopted to prevent overfitting by monitoring the loss on input/output pairs obtained following the same procedure with $\mathcal{D}^{\text{Val}}$. The training was conducted for $10,000$ epochs on a single NVIDIA V100 32GB GPU at the Jean-Zay CNRS supercomputer.

4 Results

4.1 Ablation Study on Adversarial Loss Term

This section reports results from the ablation study on hyperparameter α, which controls the weight of the adversarial loss component in Eq. 8. This investigation focuses on understanding how adversarial training affects the trade-off between reconstruction fidelity and distributional alignment in the proposed hybrid architecture.

4.1.1 Evaluation metrics

Two metrics are evaluated on the test set $\mathcal{D}^{\text{Test}}$ to assess model performance. Reconstruction accuracy is measured using Normalized Root Mean Square Error (NRMSE). For each test case i in $\mathcal{D}^{\text{Test}}$ and velocity component c, the NRMSE is defined as:

$$\text{NRMSE}_c^{(i)} = \frac{\sqrt{\frac{1}{N_t N_x N_y} \sum_{j,k,p} \left(S'^{(i)}_{j,k,p,c} - \tilde{S}'^{(i)}_{j,k,p,c} \right)^2}}{\max_{j,k,p}(S'^{(i)}_{j,k,p,c}) - \min_{j,k,p}(S'^{(i)}_{j,k,p,c})} \tag{11}$$

where $S'^{(i)}_{j,k,p,c}$ represents fluctuation velocity with mean $\mu_c^{(i)}$ (defined in Eq. 2) subtracted, and j, k, p are the time, x-direction and y-direction indices, respectively.

Distributional alignment is quantified using the Wasserstein distance, which measures the minimum cost of transforming one distribution into another, making it robust for comparing complex flow field distributions. Smaller values indicate better alignment between generated and true distributions. For each test case i and velocity component c, the Wasserstein distance is:

$$W_c^{(i)} = \inf_{\gamma \in \Pi\left(p(\mathbf{S}_c^{(i)}),\, p(\tilde{\mathbf{S}}_c^{(i)})\right)} \left(\mathbb{E}_{(\mathbf{S}_c^{(i)}, \tilde{\mathbf{S}}_c^{(i)}) \sim \gamma} \left[\| \mathbf{S}_c^{(i)} - \tilde{\mathbf{S}}_c^{(i)} \| \right] \right) \tag{12}$$

where $\|\cdot\|$ is the Euclidean norm, and $p(\mathbf{S}_c^{(i)})$, $p(\tilde{\mathbf{S}}_c^{(i)})$ are the empirical distributions of true and reconstructed velocity fields, respectively. The infimum is taken over all joint distributions $\gamma \in \Pi(p(\mathbf{S}_c^{(i)}), p(\tilde{\mathbf{S}}_c^{(i)}))$ with the specified marginals.

4.1.2 Results analysis

Table 2 reports the ablation study results for both velocity components, revealing varied impacts of α. Percentages indicate changes relative to the benchmark model, where negative values represent improvements and positive values represent degradation. For brevity, the i index denoting test cases is dropped, with NRMSE_u and W_u denoting respective metrics for the u-component, and similarly for the v-component.

Table 2: Results of the ablation study on the hyperparameter α, controlling the weight of the adversarial loss component (Eq. 8).

Metric	Case	Benchmark	α=0.2	α=0.02	α=0.002
NRMSE_u	Initial	0.0621	0.0618 (-0.48%)	0.0629 (+1.29%)	0.0627 (+0.97%)
	Upper	0.0927	0.0926 (-0.11%)	0.0929 (+0.22%)	0.0925 (-0.22%)
	Transition	0.1088	0.0993 (-8.64%)	0.1090 (+0.18%)	0.1087 (-0.09%)
	Lower	0.0842	0.0837 (-0.60%)	0.0843 (+0.12%)	0.0843 (+0.12%)
	Asynchrony	0.0752	0.0752 (+0.00%)	0.0760 (+1.06%)	0.0758 (+0.80%)
NRMSE_v	Initial	0.0935	0.0928 (-0.75%)	0.0954 (+2.03%)	0.0961 (+2.78%)
	Upper	0.1061	0.1044 (-1.60%)	0.1058 (-0.28%)	0.1053 (-0.75%)
	Transition	0.1151	0.0968 (-15.90%)	0.1170 (+1.65%)	0.1158 (+0.61%)
	Lower	0.0898	0.0880 (-2.00%)	0.0910 (+1.34%)	0.0900 (+0.22%)
	Asynchrony	0.0988	0.0986 (-0.20%)	0.0996 (+0.81%)	0.1003 (+1.52%)
W_u	Initial	0.0027	0.0029 (+7.41%)	0.0028 (+3.70%)	0.0031 (+14.81%)
	Upper	0.0036	0.0044 (+22.22%)	0.0038 (+5.56%)	0.0046 (+27.78%)
	Transition	0.0058	0.0060 (+3.45%)	0.0058 (+0.00%)	0.0068 (+17.24%)
	Lower	0.0045	0.0047 (+4.44%)	0.0041 (-8.89%)	0.0049 (+8.89%)
	Asynchrony	0.0087	0.0087 (+0.00%)	0.0080 (-8.05%)	0.0094 (+8.05%)
W_v	Initial	0.0026	0.0022 (-15.38%)	0.0019 (-26.92%)	0.0023 (-11.54%)
	Upper	0.0056	0.0048 (-14.29%)	0.0042 (-25.00%)	0.0051 (-8.93%)
	Transition	0.0072	0.0049 (-31.94%)	0.0060 (-16.67%)	0.0069 (-4.17%)
	Lower	0.0063	0.0048 (-23.81%)	0.0048 (-23.81%)	0.0061 (-3.17%)
	Asynchrony	0.0087	0.0078 (-10.34%)	0.0065 (-25.29%)	0.0083 (-4.60%)

For NRMSE_u, the $\alpha = 0.2$ configuration shows a 8.64% improvement in the transition regime and smaller gains in initial (0.48%) and lower (0.60%) regimes. Conversely, the other α values produce minimal changes, with $\alpha = 0.002$ showing small degradations in some regimes (e.g., 1.29% increase for the initial). NRMSE_v follows similar trends, with $\alpha = 0.2$ achieving the most substantial improvement of 15.90% in the transition regime and moderate gains in upper (1.60%) and initial (0.75%) regimes. Lower α values again yield less favorable results. The $\alpha = 0.2$ configuration emerges as most effective for reconstruction accuracy, particularly excelling in the Transition regime.

Wasserstein distance results reveal asymmetric performance between velocity components. For W_u, $\alpha = 0.2$ unexpectedly degrades performance in the upper branch (22.22% increase) while showing minor increases for the other regimes. In contrast, $\alpha = 0.02$ provides improvements in lower (8.89% reduction) and asynchrony (8.05% reduction) regimes. The $\alpha = 0.002$ configuration consistently increases W_u values, indicating minimal adversarial weighting offers little benefit. W_v results present a more favorable picture for β-VAE-GAN models. All α configurations show improvements across most cases, with $\alpha = 0.02$ achieving the largest gains: 26.92% in initial, 25.00% in upper, and 25.29% in asynchrony. The $\alpha = 0.2$ model also delivers strong improvements, including a substantial 31.94% reduction in the transition case. Even $\alpha = 0.002$ provides consistent but modest improvements.

Overall, the adversarial term seems to improve the distributional alignment of the cross-stream velocity component v, sometimes at the expense of worse u distribution alignment. The $\alpha = 0.2$ configuration provides the optimal balance, delivering better reconstruction accuracy while maintaining reasonable distributional performance.

4.2 Latent-Space Trajectory Prediction

Having established $\alpha = 0.2$ as the optimal adversarial loss configuration in the preceding analysis, this section examines the bidirectional transformer's capability to map cylinder displacement $\mathbf{y}_{\text{cyl}}$ to target latent-space trajectories $\boldsymbol{\zeta}$ encoded by the β-VAE-GAN.

Two cases from the test subset (Section 2.2) are analyzed in detail: the upper branch ($U^* = 5.56$) and the lower branch ($U^* = 8.03$). These operating conditions are typical of the extended synchronization region characterizing low-mass VIV systems. They were selected because, although both cases lie within the resonance region, they exhibit different dynamics. Specifically, the lower branch case is characterized by highly periodic oscillations, while the upper branch exhibits an irregular chaotic dynamics [40, 41].

4.2.1 Methodological Framework

All analyses use normalized representations to facilitate quantitative comparison. The predicted and target latent variables are normalized by the maximum magnitude of the target latent variables according to:

$$\zeta_{j,i}^{\text{norm}} = \zeta_{j,i} / \max_j \left(\sqrt{(\zeta_{j,1}^{\text{gt}})^2 + (\zeta_{j,2}^{\text{gt}})^2 + (\zeta_{j,3}^{\text{gt}})^2} \right) \tag{13}$$

where $\zeta_{j,i}$ is the i-th component of the latent vector at time step t_j, and $\zeta_{j,i}^{\text{gt}}$ represents the target trajectory obtained from β-VAE-GAN encoding of PIV flow field snapshots.

4.2.2 Temporal Evolution and Statistical Properties

The different dynamical behaviors of the two operating conditions are well captured in the temporal and statistical characterization of the latent trajectories (Figure 7a and Figure 7b).

The temporal evolution of all three latent variables in the upper regime (Figure 7a, left panels) is characterized by pronounced non-harmonic features. This is most evident in ζ_1, but also observed in the strong modulation of ζ_2 and ζ_3, which present more quasi-periodic behavior.

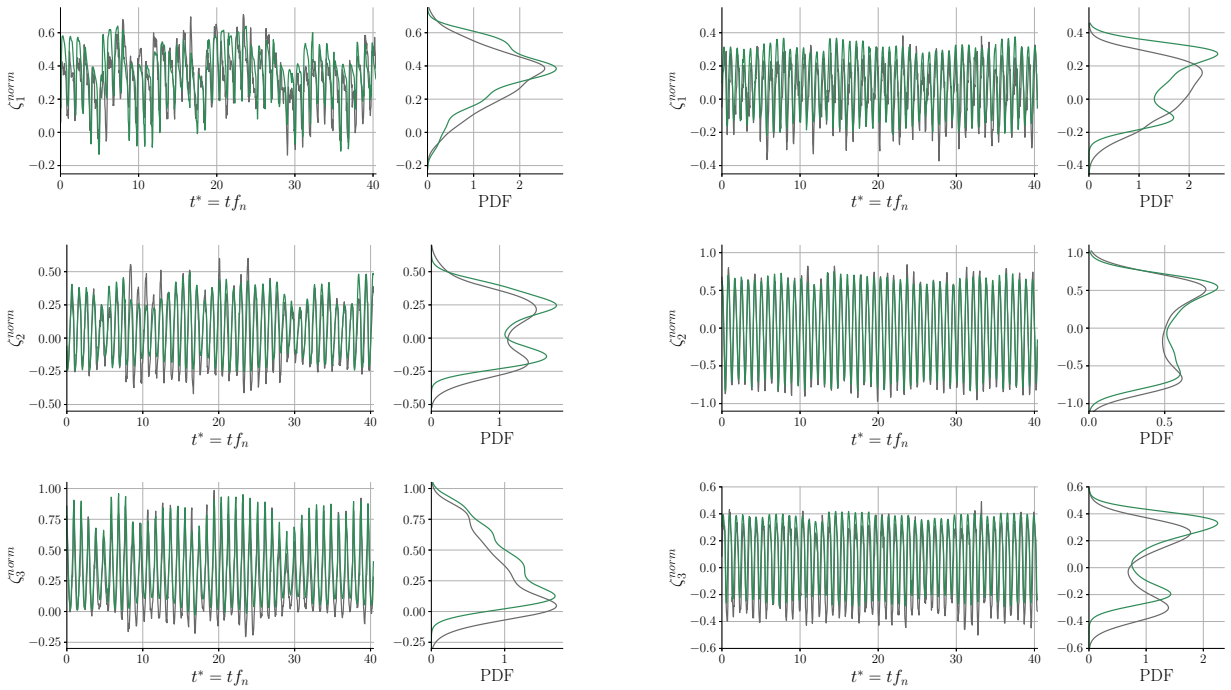
The transformer model predictions demonstrate adequate phase synchronization across all dimensions with a mean absolute percentage error of 8.87%. Discrepancies in amplitude reconstruction are evident, with the model generally underestimating peak magnitudes observed in the target trajectory.

The probability density function (PDF) analysis (Figure 7a, right panels) reveals distinct shapes for each latent variable. The ζ_2 component manifests a bimodal distribution with approximately symmetric peaks, which is characteristic of dynamics with a dominant sinusoidal component. In contrast, ζ_1 exhibits significant positive skewness (right-tail asymmetry), while ζ_3 demonstrates pronounced negative skewness (left-tail asymmetry). These complementary asymmetric distributions reflect the observed phase-modulated oscillatory dynamics of this regime, consistent with the irregular vortex dynamics observed experimentally by Khalak and Williamson [40].

The transformer predictions demonstrate effective shape preservation; however, they also manifest systematically narrower PDFs. This variance deficit indicates a limitation in the transformer's capacity to fully capture the variability of the test subset, particularly affecting representation of non-Gaussian statistical features as the asymmetric tails in ζ_1 and ζ_3.

The lower-branch latent variables evolution (Figure 7b, left panels) exhibits modifications in the flow dynamics. While the upper and lower branches are well characterised by pairs of counter-rotating vortices, the wake dynamics is different. Increased regularity is observed across all latent variables, especially ζ_2 and ζ_3, which present reduced amplitude modulations. This weakly modulated nature is directly related to the stable 2P shedding observed for this regime [40], which represents a more coherent, phase-stabilized flow organization. Despite this regime transition, the transformer maintains robust phase synchronization across all dimensions with a mean absolute percentage error of 9.92%. Amplitude discrepancies persist and, in some cases, increase moderately compared to the upper-branch case; this is most notable for ζ_1, where the attenuation of predicted amplitudes relative to the target trajectory is most pronounced.

The flow reorganization is most evident in the lower branch latent variable distributions (Figure 7b, right panels). Compared to the upper branch, ζ_2 and ζ_3 display statistical homogenization, both exhibiting similar bimodal PDFs. These distributions are characteristic of the two dominant oscillation modes, consistent with the phase portrait shown in the following section analysis (Figure 9d) and the high vortex periodicity observed by Khalak and Williamson [40]. The



(a) (b)

Figure 7: Temporal evolution comparison between target (gray) and predicted (green) latent variables for (a) upper branch ($U^* = 5.56$) and (b) lower branch ($U^* = 8.03$). Time is normalised by the cylinder natural oscillation frequency ($t^* = tf_n$). Each panel presents time series data (left) and corresponding probability density functions (right).

ζ_1 PDF also exhibits homogenization, with considerably reduced skewness compared to the upper branch, indicating weakened non-harmonic phenomena.

The transformer successfully reproduces the qualitative bimodal structure of ζ_2 and ζ_3 PDFs. However, the predicted ζ_1 distribution exhibits bimodal behavior not observed in the target latents, suggesting model limitations in capturing the statistical properties of this component. As in the upper-branch case, reduced variance is observed across all predictions.

4.2.3 Phase Portraits and Attractor Topology

Phase-space portraits are used to assess further the transformer's predictive performance in characterizing the invariant manifolds that govern system behavior. Figures 8-9 present comprehensive visualizations of the latent-space phase portraits for both flow regimes. The topological correspondence between predicted and target attractors is quantitatively assessed through a convex hull analysis. The convex hull $\mathcal{H}(P)$ of phase-portrait points $P = \{p_1, p_2, ..., p_n\}$ is defined as the minimal convex enclosure containing all points in P:

$$\mathcal{H}(P) = \left\{ \sum_{i=1}^{n} \gamma_i p_i \mid \gamma_i \geq 0, \sum_{i=1}^{n} \gamma_i = 1 \right\}. \tag{14}$$

The computations use Delaunay triangulation to decompose $\mathcal{H}(P)$ into simplices, with point containment determined via barycentric coordinate analysis. Volume ratios $V_{\text{pred}}/V_{\text{gt}}$ and containment analysis then quantify state space volume and exploration, where containment analysis measures the fraction of predicted trajectory points that lie within the target attractor's convex hull and the fraction of target points within the predicted convex hull.

The upper branch regime (Figure 8) presents a roughly annular structure with points broadly spread in the radial and axial direction. This ring-like structure is most evident in the ζ_2-ζ_3 projection, where quasi-periodic trajectories are observed. This projection shape is comparable to the one observed in the two dominant POD modes of the same branch found in literature, for example, in Janocha *et al.* [42] and Riches *et al* [43]. This topological feature also aligns with the bimodal PDF of ζ_2 observed in the temporal evolution analysis (Figure 7a). These observations suggest that these two

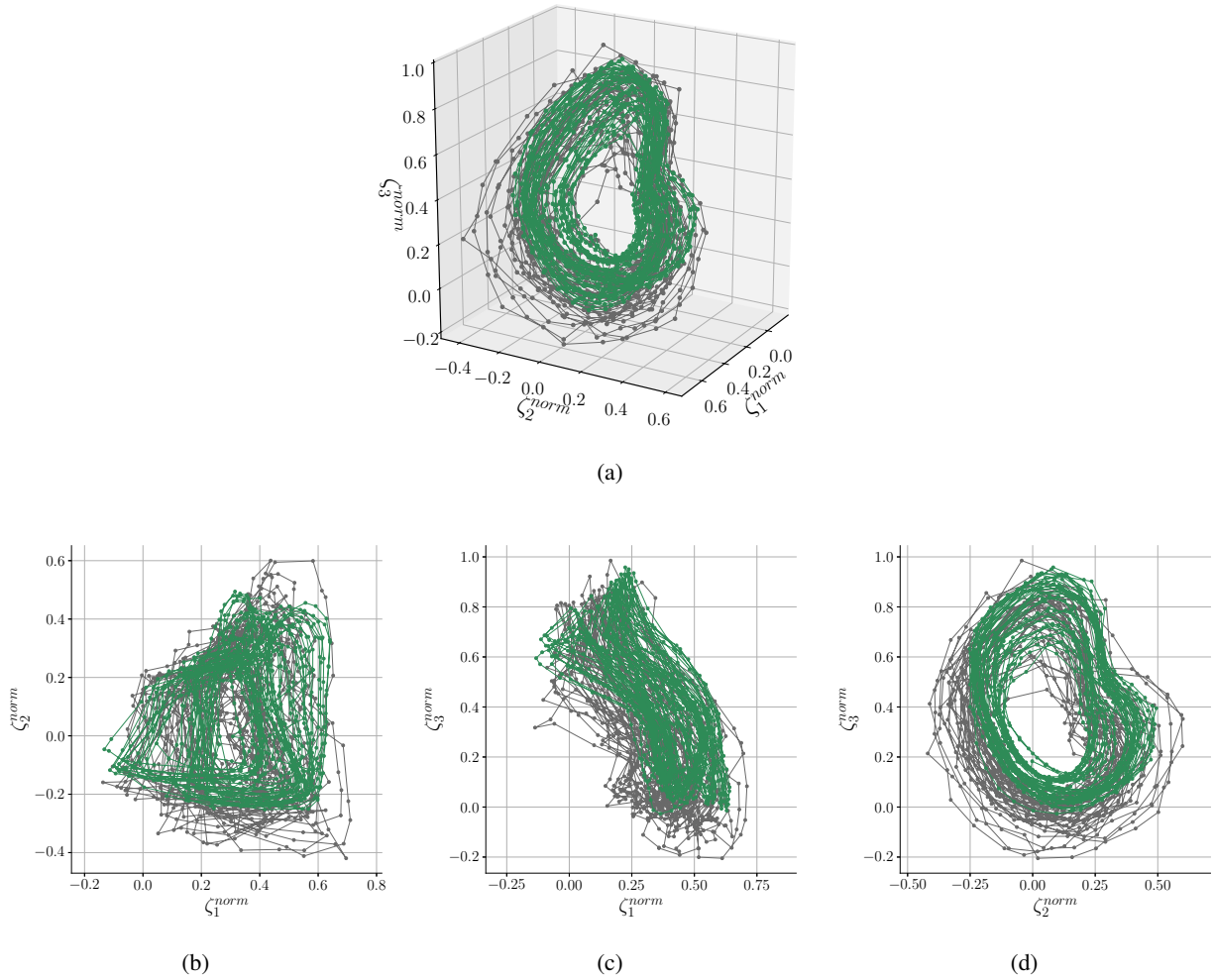


Figure 8: Phase portraits comparing target (gray) and predicted (green) latent-space trajectories of the upper branch regime ($U^* = 5.56$). (a) Three-dimensional representation. (b) ζ_1-ζ_2 projection. (c) ζ_1-ζ_3 projection. (d) ζ_2-ζ_3 projection.

latent dimensions collectively encode the principal oscillatory mechanisms governing vortex formation and shedding. The ζ_1-ζ_2 projection, instead, exhibits a diffuse, irregular distribution with concentrated density in the central region (Figure 8b). The pattern reflects the non-harmonic features observed in the temporal evolution of these variables, and it is comparable to the third POD mode pair found in the analysis of Riches *et al* [43], where the authors associate it with an intermittent behavior of the wake. The ζ_1-ζ_3 projection displays a linear relation with negative slope (Figure 8c) representing a phase opposition relationship between these two modes, in accordance with the strong anti-correlation $\rho(\zeta_1, \zeta_3) \approx -0.66$ quantified in the correlation analysis (Section 4.2.4). This pattern is consistent with the oppositely directed skewness observed in their respective probability density functions (Figure 7a) and is not evident in linear POD analyses. The observed anti-correlation between ζ_1 and ζ_3 may relate with evidence of competing flow modes in the upper branch regime. Zhao et al. [41] reported chaos arising from competition between distinct vortex shedding modes, while Morse and Williamson [44] discovered overlapping vortex formation patterns in this regime. Specifically, Morse and Williamson identified intermittent switching between the standard 2P mode (typical of lower branch) and a newly discovered 2Po mode. This 2Po mode occurs at peak resonant response, the condition corresponding to the upper branch test case examined in this analysis. The strong anti-correlation, asymmetric PDF distributions, and linear phase relationship collectively suggest that ζ_1 and ζ_3 may capture these competing dynamical states. The intermittent behavior previously associated with ζ_1-ζ_2 projection further supports this interpretation, potentially reflecting the mode-switching dynamics observed by Morse and Williamson. Therefore, this may represents a form of nonlinear mode interaction that

the ML approach captures while linear decomposition methods cannot detect. However, further investigation would be needed to establish the definitive physical significance of this relationship.

The transformer predictions align with the target attractor shape but exhibit volumetric contraction. Volume ratio analysis indicates the predicted attractor occupies 55% of the target manifold volume ($V_{\text{pred}}/V_{\text{gt}} = 0.55$), consistent with the peak underestimation observed in temporal evolution analysis. Containment analysis reveals that 98.4% of predicted attractor points lie within the target attractor hull, while only 51.4% of target points are contained within the predicted hull. This asymmetric containment pattern indicates the model successfully captures core attractor structure but fails to reproduce the full extent of system variability, as also observed in the PDF analysis.

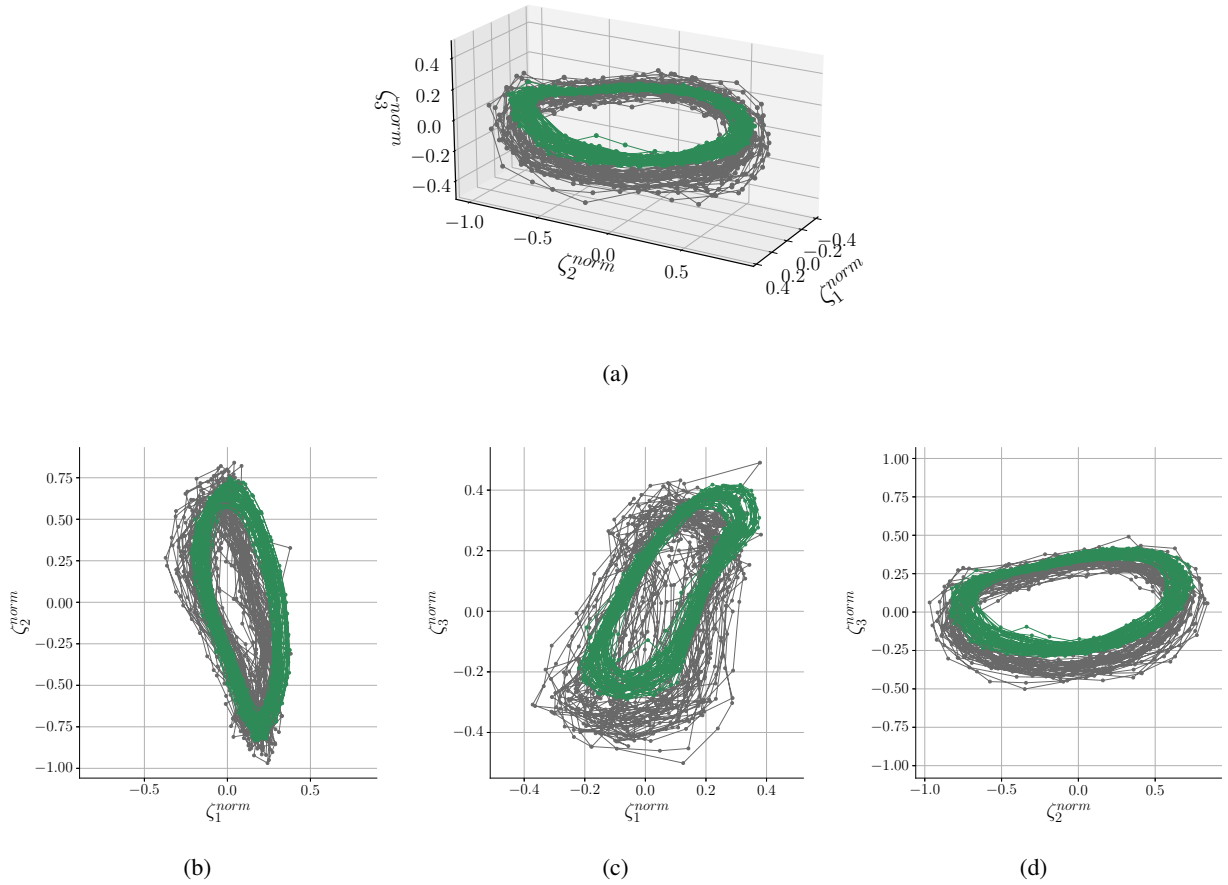


Figure 9: Phase portraits comparing target (gray) and predicted (green) latent-space trajectories of the lower branch regime ($U^* = 8.03$). (a) Three-dimensional representation. (b) ζ_1-ζ_2 projection. (c) ζ_1-ζ_3 projection. (d) ζ_2-ζ_3 projection.

The lower branch manifold (Figure 9) presents well defined annular structures. The points spread in the axial and radial directions is more constrained with respect to the upper branch regime. This ring structure is consistent with the highly periodic vortex dynamics observed by Khalak and Williamson [40] and with the increased regularity observed in the trajectories temporal evolution (Figure 7b, left). As in the upper branch, this annular shape is most evident in the ζ_2-ζ_3 projection, consistent with bimodal PDFs observed in the statistical analysis (Figure 7b, right). The projection shape is comparable to the one of the dominant POD mode pair found in Janocha *et al.* [42], who also report a reduced radial point spread of the modes compared to the upper branch, as observed in this work. As for the upper branch, these modes are linked to the principal oscillatory mechanisms. Instead, ζ_1-ζ_2 and ζ_1-ζ_3 projections appear as elongated elliptical structures oriented in opposed diagonal directions. The orientation of these structures encode specific latent variables relationship, that will be further studied in Section 4.2.4. These elliptical patterns are not evident in any of the linear decomposition study found in the literature and is therefore believed to be linked to nonlinear mode interactions captured by the ML approach, as also suggested by the irregular evolution of ζ_1 observed in the time analysis.

The transformer predictions align with the target attractor shape but exhibit greater volumetric contraction than the upper branch regime. The predicted attractor occupies only 23% of the target manifold volume ($V_{\text{pred}}/V_{\text{gt}} = 0.23$). As for the upper branch, containment analysis shows that almost all predicted points (97.5%) lie within the target space region. However, only 28.7% of target points lie within the predicted region. Therefore, consistent with the reduced variability observed in PDF analysis, the transformer predictions fail to represent the full system extent but successfully capture the core attractor.

4.2.4 Correlation Structure

Correlation analysis is performed to characterize the relationships between latent variables and assesses the transformer's capacity to reproduce these dependencies. Figure 10 presents correlation matrices for both flow regimes. Each matrix element is defined by the Pearson correlation coefficient:

$$\rho(\zeta_i, \zeta_j) = \frac{\mathbb{E}[(\zeta_i - \mu_{\zeta_i})(\zeta_j - \mu_{\zeta_j})]}{\sigma_{\zeta_i}\sigma_{\zeta_j}}, \tag{15}$$

where μ_{ζ_i} is the ensemble mean of the i-th latent variable and σ_{ζ_i} is its standard deviation.

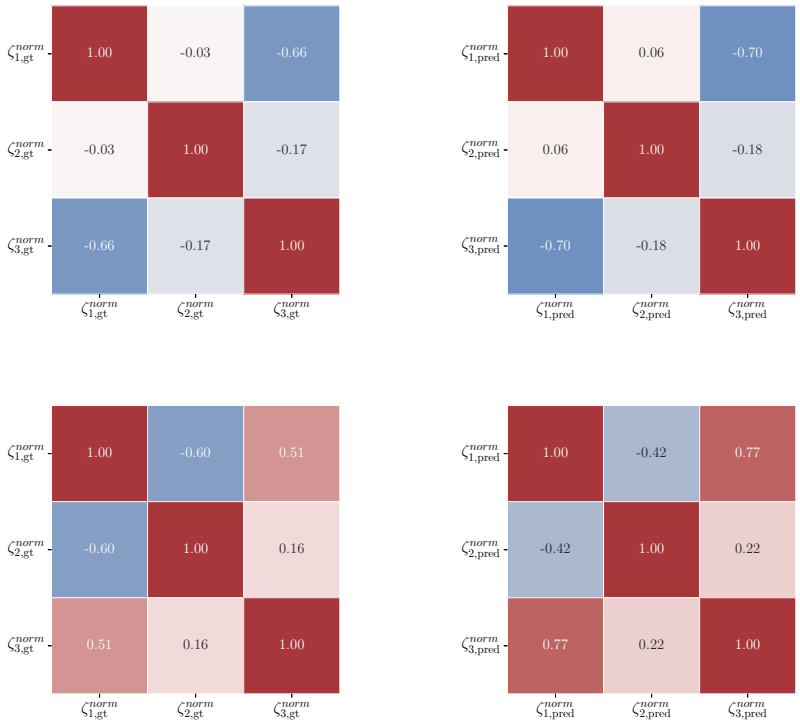


Figure 10: Left: Target correlation matrices for the latent variables. Right: Prediction correlation matrices. Top row: upper branch regime ($U^* = 5.56$). Bottom row: lower branch regime ($U^* = 8.03$).

In the upper-branch regime, the target correlation matrix (Figure 10, top left) reveals a near-diagonal structure with predominantly uncorrelated latent variables, except for a single dominant anti-correlation $\rho(\zeta_1, \zeta_3) \approx -0.66$. This structure indicates that the β-VAE-GAN encoder has disentangled the flow physics into largely independent latent features. The pronounced anti-correlation corresponds directly to the linear relationship observed in the ζ_1-ζ_3 phase projection (Figure 8c), signifying phase opposition between these complementary modes.

The transformer prediction correlation matrix closely reproduces the target structure, with most elements showing only minor quantitative departures from target values. For the ζ_1-ζ_2 pair, the correlation sign changes from negative $\rho(\zeta_1, \zeta_2) = -0.03$ to positive $\rho(\zeta_1, \zeta_2) = 0.06$. However, both correlations are negligible, confirming that these variables remain essentially uncorrelated in both cases.

The lower branch regime reveals different mode relationships. While the two dominant oscillation modes $\zeta_2 - \zeta_3$ presents correlation values $\rho(\zeta_2, \zeta_3) \approx 0.16$ comparable to the upper branch, two additional correlated pairs emerge: ζ_1-ζ_2 and ζ_1-ζ_3. These pairs exhibit opposite signed correlations, consistent with the opposing orientations of their elliptical phase portrait structures observed in Figure 9. The elliptical orbits also suggest oscillatory behavior. Under the simplified assumption of equal-amplitude sinusoidal oscillators, correlations approximately relate to phase differences as:

$$\rho(\zeta_i, \zeta_j) \approx \cos(\Delta\phi_{ij}). \tag{16}$$

Following this analogy, ζ_1-ζ_2 exhibits a phase offset of approximately $\Delta\phi_{12} \approx -53°$, while ζ_1-ζ_3 shows $\Delta\phi_{13} \approx 60°$. These nearly complementary phase relationships suggest ζ_1 may mediate coupling between the two dominant oscillatory modes. However, the assumption made may be an over-simplification of the more rich interaction between these modes and further investigation is required to understand the physical mechanism encoded by this variable.

The transformer prediction correlation matrices show larger departures from the target values than what observed in the upper branch regime. However, correlation signs remain unchanged, indicating preservation of the fundamental variables relationships. These departures are consistent with the higher prediction errors and the ζ_1 PDF mismatch discussed in the temporal and statistical analysis. The larger deviations suggest the transformer encounters greater difficulty accurately reproducing this flow regime, though it successfully captures the underlying mode coupling structure. This increased prediction challenge for the lower branch is also consistent with the greater volumetric contraction observed in phase space analysis.

4.3 Combining the transformer and the β-VAE-GAN decoder: VIVALDy inference

Having established the individual performances of the β-VAE-GAN decoder in flow field reconstruction (Section 4.1) and the bidirectional transformer in latent-space trajectory prediction (Section 4.2), this section now integrates these two models to form the complete VIVALDy inference framework and evaluates its end-to-end performance in predicting velocity fields of the test set $\mathcal{D}^{\text{Test}}$ from cylinder displacement y_{cyl}.

4.3.1 Reconstruction Accuracy and Distribution Alignment

Table 3 quantifies VIVALDy reconstruction accuracy and distributional alignment performance using the metrics introduced in Section 4.1.1. Performance is evaluated relative to a baseline using latent variables encoded from the ground truth PIV snapshots, thereby isolating the transformer's contribution to overall system performance. For brevity, NRMSE$_u$ and W_u denote the respective metrics for the u-component, with analogous notation for the v-component.

Table 3: Quantitative metrics for VIVALDy inference performance across different flow regimes. Percentages in parentheses indicate changes relative to the reference case where ground truth encoded latent variables are fed into the same decoder compared to using transformer-predicted latents. Red values highlight a significant worsening in performance.

Flow Regime	NRMSE$_u$	NRMSE$_v$	W_u	W_v
Initial	0.0960 (+55.33%)	0.1912 (+106.0%)	0.0033 (+13.72%)	0.0026 (+18.18%)
Upper	0.1023 (+10.47%)	0.1211 (+16.00%)	0.0052 (+18.18%)	0.0039 (-18.75%)
Transition	0.1180 (+18.83%)	0.1483 (+53.20%)	0.0050 (-16.66%)	0.0075 (+53.06%)
Lower	0.1042 (+24.49%)	0.1185 (+34.66%)	0.0045 (-4.255%)	0.0079 (+64.58%)
Asynchrony	0.0811 (+7.846%)	0.1204 (+22.11%)	0.0086 (-1.149%)	0.0083 (+6.410%)

NRMSE values are generally higher than the baseline but remain low across all flow regimes. The transformer-predicted latent variables introduce accuracy losses in specific regimes. The initial regime shows increases of 55.33% for NRMSE$_u$ and 106% for NRMSE$_v$, likely due to the nearly static cylinder motion in this operating condition providing a low signal-to-noise ratio for the transformer input. The transition regime exhibits increases of 18.83% for NRMSE$_u$ and 53.20% for NRMSE$_v$, which can be linked to this regime's absence from the training set, requiring greater model generalizability. Instead, the lower branch regime show increases of 24.49% for NRMSE$_u$ and 34.66% for NRMSE$_v$, which are consistent with the latent space analysis findings, where a mismatch for the ζ_1 PDF prediction and a phase space volumetric contractions were observed. Despite these increases, the model demonstrates robustness with NRMSE$_u$ values remaining below 0.12 and NRMSE$_v$ values below 0.20 across all regimes.

For Wasserstein distances, the results reveal asymmetric performance between velocity components. The u-component distributions generally show modest improvements or degradations. In contrast, the v-component exhibits distributional misalignments, particularly in the transition (+53.06%) and lower (+64.58%) regimes. This asymmetry suggests the transformer more effectively captures streamwise flow statistics, possibly due to stronger coupling between cylinder displacement and streamwise velocity fluctuations.

4.3.2 Phase Averaged Flow Visualization

Phase-averaged flow visualizations complement the quantitative NRMSE results. Various signals are used in the literature to define reference phases, such as pressure signals [45] or transverse velocity at a point in the near wake [46]. Here, phase averaging is based on the cylinder position, following O'Neill *et al.* [47]. This approach requires computing the Hilbert transform of the cylinder displacement signal to define the instantaneous phase angle. Each snapshot is then assigned to a phase bin based on its corresponding time instant, and averaged fields are computed for each bin. Following O'Neill *et al.* [47], 36 bins are used. The obtained phase-averaged fields represent the coherent velocity structures that define the dominant wake topology.

Figure 11 shows the computed phase-averaged fields for the upper and lower branch regimes analyzed in Section 4.2. Distinct wake topologies are clearly visible. The upper branch (top row) exhibits a broader wake with widely spaced vortex pairs, comparable to the 2Po vortex-shedding mode identified by Morse and Williamson [44]. Instead, the lower branch (bottom row) presents a narrower wake with more closely spaced vortex structures, closely resembling the 2P vortex-shedding mode reported by Morse and Williamson [44] and O'Neill *et al.* [47].

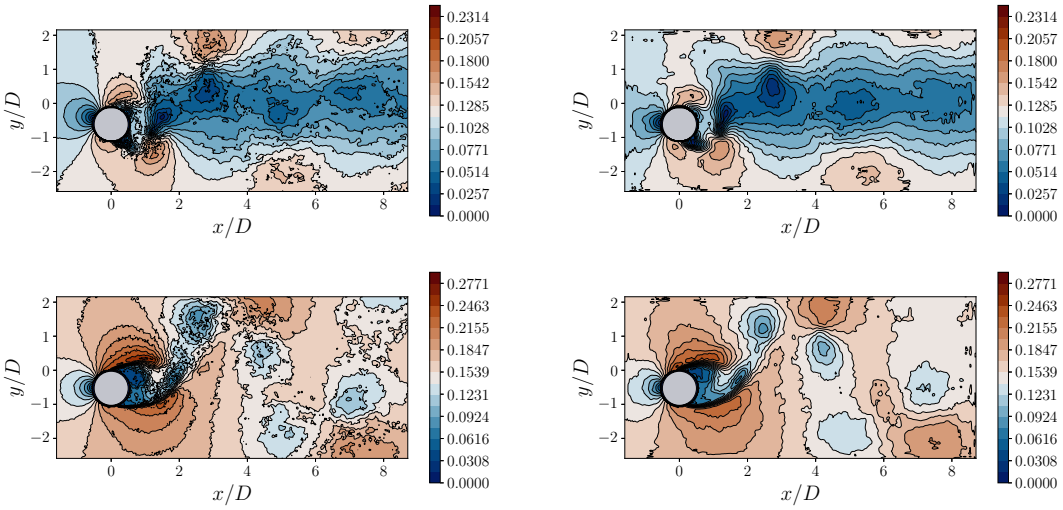


Figure 11: Phase-averaged velocity field comparison between ground truth (left) and VIVALDy predictions (right), colored by velocity magnitude $u_{\text{mag}} = \sqrt{u^2 + v^2}$. Results shown for upper branch ($U^* = 5.56$, top row) and lower branch ($U^* = 8.03$, bottom row) test cases. For each case, the depicted phase-average corresponds to the first bin.

VIVALDy's phase-averaged predictions (right column) closely align with ground truth (left column). The model reproduces both wake topologies with only minor deviations from target vortex patterns, preserving all dominant wake structures. Consequently, the NRMSE errors reported previously reflect finer flow structures, residual stochastic fluctuations not observed after the phase-averaging process, and measurement noise.

4.3.3 Probability Density Function Visualization

The flow statistical distribution visualization complements the quantitative Wasserstein distances results. Gaussian kernel density estimation was used to compute the probability density functions of the ground truth PIV snapshots and the VIVALDy predictions.

The computed PDFs for the upper and lower branch regimes are shown in Figure 12. The two operating conditions exhibit distinct velocity distributions, reflecting their different wake dynamics. Both cases show bimodal v-component distributions characteristic of alternating vortex shedding, while the u-component distributions differ significantly between regimes. The lower branch case, characterized by highly periodic oscillations, displays a symmetric u-

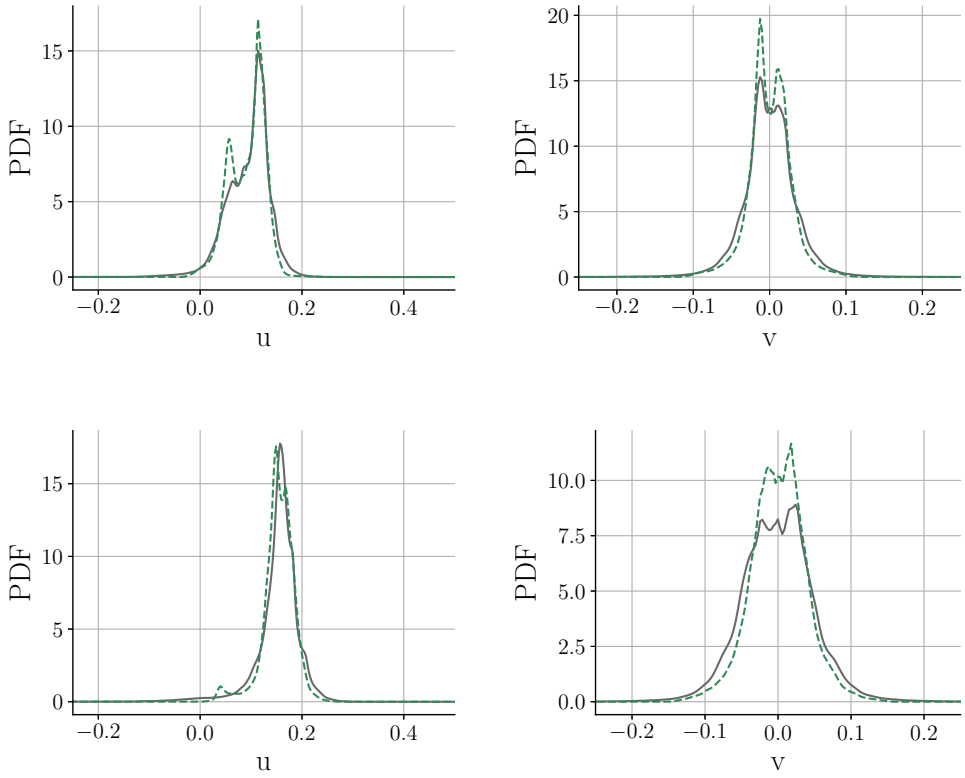


Figure 12: Probability density functions comparison between ground truth (gray) and VIVALDy predictions (green). Results shown for upper branch ($U^* = 5.56$, top row) and lower branch ($U^* = 8.03$, bottom row) test cases. For each case, the depicted phase-average corresponds to the first bin.

component distribution. In contrast, the upper branch exhibits an asymmetric u-component distribution, reflecting its irregular dynamics.

VIVALDy's predictions show close shape preservation and reasonable alignment with the ground truth PDFs of both operating conditions distribution. However, some deviation are observed. In the upper branch u-component, a significant overestimation is observed in a secondary peak. Peak overestimation and lower variance are also observed for the v-component of both regimes, consistent with the higher Wasserstein distances found for this velocity component. The predictioin asymmetry quantified in the Wasserstein distance is also observable in the distribution, with the u-component predictions showing closer agreement with the ground truth than the v-component ones.

5 Discussion

5.1 Rate-Distortion-Perception Trade-offs in Adversarial Training

The empirical results of the ablation study (Section 4.1) show that improvements in distributional alignment do not necessarily correspond to improvements in reconstruction accuracy. Indeed, the model trained with $\alpha = 0.02$ showed lower Wasserstein distances but worse NRMSE compared to the benchmark. This apparent contradiction can be understood through information theory principles. Classical rate-distortion theory quantifies the fundamental trade-off between compression efficiency (rate) and reconstruction quality (distortion) [48, 49], however it does not account for statistical fidelity. Blau and Michaeli [50] extended this framework to include distributional alignment (termed perception quality) through the rate-distortion-perception trade-off. They demonstrate that requiring high distributional alignment generally elevates the rate-distortion curve, necessitating sacrifices in either compression rate or reconstruction accuracy.

The ablation study results align with this theoretical framework. Because the latent space dimension remains fixed at three, the compression rate is constant in this work. Consequently, increased distortion is expected and empirically

observed for $\alpha = 0.02$. Conversely, with minimal adversarial weighting ($\alpha = 0.002$), slightly improved NRMSE values are observed for both velocity components, but at the cost of worse Wasserstein distances, consistent with the Blau and Michaeli framework. Notably, $\alpha = 0.2$ appears to yield a more advantageous trade-off, achieving better Wasserstein distances for v coupled with improved NRMSE for both components, despite slightly worse distributional alignment for the u component.

Interestingly, the $\alpha = 0.2$ model's greatest improvement occurs in the Transition case. As reported in Section 2.2, this case appears exclusively in the test set. This improved performance on an unseen operating condition suggests that, by forcing the optimization to preserve the statistical properties of the flow field, adversarial training enhances model generalizability. Therefore, the α parameter enables the proposed hybrid β-VAE-GAN to achieve substantial improvements over the benchmark model while maintaining robust performance across diverse operating conditions.

5.2 Reduced Variance and Model Limitations

Throughout the latent and inference analyses, VIVALDy's predictions consistently exhibit narrower PDFs and latent space volume contractions compared to the experimental fields. This prediction behavior reflects several fundamental challenges facing the model.

The chosen β-VAE-GAN architecture compress the data into just three latent variables, forcing the model to retain only dominant structures while discarding finer details. This drastic dimensionality reduction combines with two limitation imposed by the experimental dataset: limited training data and inherent measurement noise. While this work uses 900 snapshots per operating condition (10,800 total), comparable autoencoder-based studies on numerical simulations datasets typically access to 10^5 samples [24, 51]. Other studies, access to data quantities comparable to this study but operates on single flow conditions [52], whereas VIVALDy must generalize across diverse operating regimes. Given these constraints, the narrower PDFs can be interpreted as an implicit model regularization that prioritizes dominant dynamical structures, enhancing robustness at the expense of reproducing the full fluctuation spectrum. Indeed, VIVALDy demonstrates adequate accuracy across all relevant flow regimes and preserves the dominant wake topology in phase-averaged visualizations. In correlation analyses, the transformer successfully maintains variable relationships, i.e the underlying mode interactions, even as these relationships vary across flow configurations. This performance level is satisfactory for surrogate models intended for flow control or rapid design optimization applications.

6 Conclusions

Reduced-order models are frameworks designed to create numerically efficient surrogate models of turbulent flows. Traditional ROM approaches face fundamental challenges in achieving computational efficiency while maintaining accuracy across diverse operating conditions. This paper introduces VIVALDy, a novel machine-learning framework that addresses these limitations through three innovations: masked convolutions to handle complex solid-fluid interfaces; a hybrid β-VAE-GAN architecture to learn informative and statistically consistent latent representations; and a bidirectional transformer to map minimal sensor inputs to the underlying flow dynamics. When validated on a vortex-induced vibration problem, VIVALDy successfully reconstructs the turbulent flow around an oscillating cylinder using only the body's displacement as input, demonstrating robust performance even in previously unseen regimes.

To demonstrate the framework's capabilities, VIVALDy is applied to a vortex-induced vibration problem: reconstructing the turbulent flow around an oscillating cylinder using only the body displacement as input measurement. This fluid-structure interaction problem represents a challenging test case due to moving solid-fluid interface and irregularly changing wake dynamics, requiring generalization across operating regimes. The model is validated against an experimental dataset spanning diverse VIV conditions, demonstrating the general applicability of the proposed innovations.

At first, an ablation study was performed comparing three β-VAE-GAN configurations with different adversarial weight parameters α against a benchmark β-VAE. The study revealed fundamental trade-offs between reconstruction accuracy and distributional alignment, consistent with rate-distortion-perception theory. The optimal configuration ($\alpha = 0.2$) enhanced model generalizability, particularly for unseen flow conditions. Using this optimal configuration, a detailed analysis of the latent space was performed on two test-set operating conditions. The results demonstrate that the encoded latent variables retain characteristic dynamical signatures of the flow, with mode pairs tracing annular orbits associated with vortex shedding oscillations in the wake. The analysis revealed nonlinear mode interactions not captured by linear decomposition methods, with anti-correlation patterns between latent variables aligning with literature evidence of competing vortex shedding modes. This suggests VIVALDy's latent space captures physically meaningful dynamics often missed by traditional linear ROM methods. The bidirectional transformer successfully predicts these attractor shapes and variable relationships using only cylinder displacement input. When integrated with the β-VAE-GAN

decoder during inference, this approach delivers robust performance across all VIV regimes, accurately reproducing both coherent wake structures and statistical properties under diverse operating conditions.

Beyond VIV applications, the hybrid architecture and masked convolution approach provide a general framework for fluid-dynamics problems involving complex geometries. The demonstrated ability to capture nonlinear mode interactions positions this approach as a valuable tool for understanding dynamics not fully captured by traditional linear methods. The successful application of adversarial training principles suggests broader potential for incorporating distribution-preserving objectives in scientific machine-learning, particularly where statistical fidelity is as important as reconstruction accuracy. Moreover, VIVALDy's single-sensor prediction capability across different flow states makes it versatile for diverse applications: as sparse sensing reconstruction [53, 52, 54], latent space data assimilation [55], initialization of lagged observations for dynamic forecasting models such as transformers [24], and development of optimal control strategies [56, 57]. Furthermore, the learned latent space topology offers geometric interpretation of flow dynamics, enabling the development of control strategies that guide dynamics along desired trajectories directly in the latent space [58]. VIVALDy thus represents an advancement toward sophisticated reduced-order models capable of mimicking physical systems, enabling practical flow control and optimization strategies.

Acknowledgments

This work was supported by the French government program "Investissements d'Avenir" (EUR INTREE, reference ANR-18-EURE-0010 and LABEX INTERACTIFS, reference ANR-11-LABX-0017-01) and was performed using HPC resources from GENCI–IDRIS (Grant 20XX-AD011015409). RV acknowledges the financial support from ERC grant no. 2021-CoG-101043998, DEEPCONTROL. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Junlei Wang, Linfeng Geng, Lin Ding, Hongjun Zhu, and Daniil Yurchenko. The state-of-the-art review on energy harvesting from flow-induced vibrations. *Applied Energy*, 267:114902, 2020. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2020.114902. URL https://www.sciencedirect.com/science/article/pii/S0306261920304141.

[2] United Nations Environment Programme Finance Initiative. Turning the Tide: How to Finance a Sustainable Ocean Recovery. Report, United Nations Environment Programme Finance Initiative (UNEP FI), 2021. URL https://www.unepfi.org/publications/turning-the-tide/. Accessed on July 6, 2025.

[3] Michael M. Bernitsas, Kamaldev Raghavan, Y. Ben-Simon, and E. M. H. Garcia. Vivace (vortex induced vibration aquatic clean energy): A new concept in generation of clean and renewable energy from fluid flow. *Journal of Offshore Mechanics and Arctic Engineering*, 130(4):041101, 09 2008. ISSN 0892-7219. doi:10.1115/1.2957913. URL https://doi.org/10.1115/1.2957913.

[4] Agathe Schmider, Franck Kerhervé, Andreas Spohn, and Laurent Cordier. Improved viv energy harvesting with a virtual damper–spring system. *Ocean Engineering*, 293:116668, 2024.

[5] C.H.K. Williamson and R. Govardhan. Vortex-induced vibrations. *Annual Review of Fluid Mechanics*, 36: 413–455, 2004. ISSN 1545-4479. doi:https://doi.org/10.1146/annurev.fluid.36.050802.122128. URL https://www.annualreviews.org/content/journals/10.1146/annurev.fluid.36.050802.122128.

[6] Gianluigi Rozza, Giovanni Stabile, and Francesco Ballarin. *Advanced Reduced Order Methods and Applications in Computational Fluid Dynamics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022. doi:10.1137/1.9781611977257. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977257.

[7] L. Cordier and M. Bergmann. Proper Orthogonal Decomposition: an overview. In *Lecture series 2002-04, 2003-03 and 2008-01 on post-processing of experimental and numerical data*, pages 1–46. Von Kármán Institute for Fluid Dynamics, 2008. ISBN 978-2-930389-80-X.

[8] L. Cordier and M. Bergmann. Two typical applications of POD: coherent structures eduction and reduced order modelling. In *Lecture series 2002-04, 2003-03 and 2008-01 on post-processing of experimental and numerical data*, pages 1–60. Von Kármán Institute for Fluid Dynamics, 2008. ISBN 978-2-930389-80-X.

[9] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[10] Boris Kramer, Benjamin Peherstorfer, and Karen E. Willcox. Learning nonlinear reduced models from data with operator inference. *Annual Review of Fluid Mechanics*, 56(Volume 56, 2024):521–548, 2024. ISSN 1545-4479. doi:https://doi.org/10.1146/annurev-fluid-121021-025220. URL https://www.annualreviews.org/content/journals/10.1146/annurev-fluid-121021-025220.

[11] Shane A McQuarrie, Cheng Huang, and Karen E Willcox. Data-driven reduced-order models via regularised operator inference for a single-injector combustion process. *Journal of the Royal Society of New Zealand*, 51(2): 194–211, 2021.

[12] Benjamin Peherstorfer and Karen Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306:196–215, 2016.

[13] Boris Kramer, Benjamin Peherstorfer, and Karen E Willcox. Learning nonlinear reduced models from data with operator inference. *Annual Review of Fluid Mechanics*, 56(1):521–548, 2024.

[14] Peter Benner, Serkan Gugercin, and Karen Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM review*, 57(4):483–531, 2015.

[15] Shane A McQuarrie, Parisa Khodabakhshi, and Karen E Willcox. Nonintrusive reduced-order models for parametric partial differential equations via data-driven operator inference. *SIAM Journal on Scientific Computing*, 45(4):A1917–A1946, 2023.

[16] Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52(Volume 52, 2020):477–508, 2020. ISSN 1545-4479. doi:https://doi.org/10.1146/annurev-fluid-010719-060214. URL https://www.annualreviews.org/content/journals/10.1146/annurev-fluid-010719-060214.

[17] Ricardo Vinuesa and Steven L Brunton. Enhancing computational fluid dynamics with machine learning. *Nature Computational Science*, 2(6):358–366, 2022.

[18] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[19] Lionel Agostini. Exploration and prediction of fluid dynamical systems using auto-encoder technology. *Physics of Fluids*, 32(6), 2020.

[20] Eurika Kaiser, Bernd R Noack, Laurent Cordier, Andreas Spohn, Marc Segond, Markus Abel, Guillaume Daviller, Jan Östh, Siniša Krajnović, and Robert K Niven. Cluster-based reduced-order modelling of a mixing layer. *Journal of Fluid Mechanics*, 754:365–414, 2014.

[21] Hamidreza Eivazi, Soledad Le Clainche, Sergio Hoyas, and Ricardo Vinuesa. Towards extraction of orthogonal and parsimonious non-linear modes from turbulent flows. *Expert Systems with Applications*, 202:117038, 2022.

[22] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[23] Yuning Wang, Alberto Solera-Rico, Carlos Sanmiguel Vila, and Ricardo Vinuesa. Towards optimal β-variational autoencoders combined with transformers for reduced-order modelling of turbulent flows. *International Journal of Heat and Fluid Flow*, 105:109254, 2024. ISSN 0142-727X. doi:https://doi.org/10.1016/j.ijheatfluidflow.2023.109254. URL https://www.sciencedirect.com/science/article/pii/S0142727X23001534.

[24] Alberto Solera-Rico, Carlos Sanmiguel Vila, Miguel Gómez-López, Yuning Wang, Abdulrahman Almashjary, Scott TM Dawson, and Ricardo Vinuesa. β-variational autoencoders and transformers for reduced-order modelling of fluid flows. *Nature Communications*, 15(1):1361, 2024.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[28] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. *Advances in neural information processing systems*, 31, 2018.

[29] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

[30] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020.

[31] R Govardhan and CHK1789246 Williamson. Modes of vortex formation and frequency response of a freely vibrating cylinder. *Journal of Fluid Mechanics*, 420:85–130, 2000.

[32] Atul Kumar Soti, Mark C Thompson, John Sheridan, and Rajneesh Bhardwaj. Harnessing electrical power from vortex-induced vibration of a circular cylinder. *Journal of Fluids and Structures*, 70:360–373, 2017.

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[34] Stephen B Pope. Turbulent flows. *Measurement Science and Technology*, 12(11):2020–2021, 2001.

[35] Francisco Huera-Huarte. Vortex-induced vibration of flexible cylinders in cross-flow. *Annual Review of Fluid Mechanics*, 57, 2025.

[36] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[37] Christian Jacobsen and Karthik Duraisamy. Disentangling generative factors of physical fields using variational autoencoders. *Frontiers in Physics*, 10:890910, 2022.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[39] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

[40] Asif Khalak and Charles HK Williamson. Motions, forces and mode transitions in vortex-induced vibrations at low mass-damping. *Journal of fluids and Structures*, 13(7-8):813–851, 1999.

[41] Jisheng Zhao, Justin S Leontini, D Lo Jacono, and John Sheridan. Chaotic vortex induced vibrations. *Physics of Fluids*, 26(12), 2014.

[42] Marek Jan Janocha, Muk Chen Ong, and Guang Yin. Large eddy simulations and modal decomposition analysis of flow past a cylinder subject to flow-induced vibration. *Physics of Fluids*, 34(4), 2022.

[43] Graham Riches, Robert Martinuzzi, and Chris Morton. Proper orthogonal decomposition analysis of a circular cylinder undergoing vortex-induced vibrations. *Physics of Fluids*, 30(10), 2018.

[44] TL Morse and CHK Williamson. Prediction of vortex-induced vibration response by employing controlled motion. *Journal of Fluid Mechanics*, 634:5–39, 2009.

[45] Rodolphe Perrin, Marianna Braza, Emmanuel Cid, Sebastien Cazin, Arnaud Barthet, Alain Sevrain, C Mockett, and F Thiele. Obtaining phase averaged turbulence properties in the near wake of a circular cylinder at high reynolds number using pod. *Experiments in Fluids*, 43(2):341–355, 2007.

[46] N Cagney and S Balabani. Wake modes of a cylinder undergoing free streamwise vortex-induced vibrations. *Journal of Fluids and Structures*, 38:127–145, 2013.

[47] Christopher M O'Neill, Graham Riches, and Chris Morton. Wake dynamics and heuristic modelling in the desynchronization region of 1-dof viv. *International Journal of Heat and Fluid Flow*, 88:108729, 2021.

[48] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4 (142-163):1, 1959.

[49] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 2006.

[50] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[51] Kai Fukami and Kunihiko Taira. Grasping extreme aerodynamics on a low-dimensional manifold. *Nature Communications*, 14(1):6480, 2023.

[52] Jan P Williams, Olivia Zahn, and J Nathan Kutz. Sensing with shallow recurrent decoder networks. *arXiv preprint arXiv:2301.12011*, 2023.

[53] Kai Fukami, Romit Maulik, Nesar Ramachandra, Koji Fukagata, and Kunihiko Taira. Global field reconstruction from sparse sensors with voronoi tessellation-assisted deep learning. *Nature Machine Intelligence*, 3(11):945–951, 2021.

[54] Abhijeet Vishwasrao, Sai Bharath Chandra Gutha, Andres Cremades, Klas Wijk, Aakash Patil, Catherine Gorle, Beverley J McKeon, Hossein Azizpour, and Ricardo Vinuesa. Diff-sport: Diffusion-based sensor placement optimization and reconstruction of turbulent flows in urban environments. *arXiv preprint arXiv:2506.00214*, 2025.

[55] Maddalena Amendola, Rossella Arcucci, Laetitia Mottet, Cesar Quilodran Casas, Shiwei Fan, Christopher Pain, Paul Linden, and Yi-Ke Guo. Data assimilation in the latent space of a neural network, 2020. URL `https://arxiv.org/abs/2012.12056`.

[56] Andrés Cremades, Sergio Hoyas, Rahul Deshpande, Pedro Quintero, Martin Lellep, Will Junghoon Lee, Jason P Monty, Nicholas Hutchins, Moritz Linkmann, Ivan Marusic, et al. Identifying regions of importance in wall-bounded turbulence through explainable deep learning. *Nature Communications*, 15(1):3864, 2024.

[57] Miguel Beneitez, Andres Cremades, Luca Guastoni, and Ricardo Vinuesa. Improving turbulence control through explainable deep learning. *arXiv preprint arXiv:2504.02354*, 2025.

[58] Kai Fukami, Hiroya Nakao, and Kunihiko Taira. Data-driven transient lift attenuation for extreme vortex gust–airfoil interactions. *Journal of Fluid Mechanics*, 992:A17, 2024.