Latent Traversals in Generative Models as Potential Flows

Yue Song[1 2] **Andy Keller**[2] **Nicu Sebe**[1] **Max Welling**[2]

Abstract

Despite the significant recent progress in deep generative models, the underlying structure of their latent spaces is still poorly understood, thereby making the task of performing semantically meaningful latent traversals an open research challenge. Most prior work has aimed to solve this challenge by modeling latent structures linearly, and finding corresponding linear directions which result in 'disentangled' generations. In this work, we instead propose to model latent structures with a learned dynamic potential landscape, thereby performing latent traversals as the flow of samples down the landscape's gradient. Inspired by physics, optimal transport, and neuroscience, these potential landscapes are learned as physically realistic partial differential equations, thereby allowing them to flexibly vary over both space and time. To achieve disentanglement, multiple potentials are learned simultaneously, and are constrained by a classifier to be distinct and semantically self-consistent. Experimentally, we demonstrate that our method achieves both more qualitatively and quantitatively disentangled trajectories than state-of-the-art baselines. Further, we demonstrate that our method can be integrated as a regularization term during training, thereby acting as an inductive bias towards the learning of structured representations, ultimately improving model likelihood on similarly structured data. Code is available at `https://github.com/KingJamesSong/PDETraversal`.

1. Introduction

Generative models such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Auto-Encoders (VAEs) (Kingma & Welling, 2014) have latent spaces that are rich in semantics, whereby traversing latent codes according to carefully chosen trajectories has the possibility to lead to semantically meaningful transformations in the generated images. However, without a carefully structured latent space, it is impossible a priori to know how to precisely construct such trajectories. A significant research effort has thus emerged to develop methods that are able to discover semantically meaningful, self-consistent, and disentangled trajectories in the latent space of pre-trained generative models. Such traversals would allow for a more controlled generation of images without needing to alter or constrain the training process of the generative model itself. Most straightforwardly, an early set of these approaches aimed to identify fixed linear directions in latent space and evolve samples along the discovered directions to create trajectories (Härkönen et al., 2020; Voynov & Babenko, 2020; Shen & Zhou, 2021). Such efforts developed valuable techniques for unsupervised learning of interpretable traversal directions but were ultimately limited by their assumption that semantics were structured linearly in latent space, and thus were prone to yielding less semantically disentangled traversals. More recently, Tzelepis et al. (2021) proposed to model nonlinear latent traversals using gradients of learned Gaussian Radial Basis Functions (RBFs) to effectively 'warp' the latent space and thereby drive latent traversals. This integrated non-linearity was demonstrated to improve the modeling of the semantic structure but again was limited by its relatively fixed shape and its static nature over the time-length of the traversal.

In this work, we introduce a more general framework which encompasses this prior work while simultaneously allowing for a significantly more flexible learned latent structure. Our approach is motivated by intuitions from physics, optimal transport, and neuroscience, and proposes to model latent traversals as the flow of particles down the gradient of a latent potential landscape. The challenge of learning a set of disentangled latent traversals then equates to the problem of learning a set of equivalent disentangled potential functions which match the semantic structure of the underlying data manifold. Traversals can then be generated by evolving samples through time following the gradient of these learned potentials. Importantly, in contrast with prior work, our framework defines the learned potential functions as physically realistic Partial Differential Equations (PDEs), thereby allowing them to vary over both time and space,

[1]University of Trento, Italy [2]University of Amsterdam, the Netherlands. Correspondence to: Yue Song <yue.song@unitn.it>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

enabling sufficiently greater flexibility of traversal paths than existing counterparts. In practice, we show that our framework can be applied to multiple different generative models under different experimental settings, and successfully improves performance on a variety of fronts. For example, with pre-trained GANs and VAEs, our framework identifies latent trajectories which are qualitatively more disentangled, and score higher on objective disentanglement metrics than state-of-the-art linear and RBF counterparts. Further, when the desired factors of variation are known a priori, our method can also be integrated into the training process of generative models by performing "supervised" latent traversals, thereby simultaneously structuring the latent space and providing users with learned latent traversal directions. We show that such integrated structures serve as a beneficial inductive bias for similarly smooth structured input transformations, and thereby improve the likelihood of structured data under the model. Moreover, our latent operator could induce the model with approximate transformation equivarience. Finally, we perform an empirical analysis of our method, demonstrating that our framework can model unambiguous traversal paths in diverse shapes. We conclude with a discussion about how many different well-known 'special' PDEs may be used to model the sample evolution, and how previous linear traversal approaches may be seen as special cases of our method.

2. Motivation

In this section, we outline the diverse set of motivations which provide useful intuition for the success of our method, in addition to outlining clear paths for potential future work.

2.1. Fluid Mechanics as Optimal Transport

Optimal Transport (OT) can be described at a high level as finding a map which moves the probability mass between a source and target distribution with minimal cost. Intuitively, this has a strong connection with latent traversals which can similarly be seen as attempting to move samples from a source probability distribution to a target probability distribution most efficiently while staying on the data manifold. For example, consider aiming to perform a traversal which changes the length of an individual's hair while leaving the rest of their traits unaffected. With the constraint that the traversal must stay on the data manifold, the most efficient traversal would not involve the transformation of multiple variables, as this would require the movement of additional mass, but instead only transform the latent code in a direction which corresponds to the transformation of a single generative factor. In essence, if we were able to learn the underlying structure of the data manifold with respect to various semantic attributes, optimal transport would give us a direct solution to how to perform disentangled traversals.

One method for solving optimal transport problems involves casting them to a fluid mechanical system (Benamou & Brenier, 2000), and solving the associated system numerically. More formally, given the source and target density functions $\rho_0(\boldsymbol{x}), \rho_T(\boldsymbol{x}) \geq 0$, if we construct a dynamical system defined by a continuous density field $\rho(\boldsymbol{x}, t) \geq 0$ and a velocity field $v(\boldsymbol{x}, t)$, where $\rho(\boldsymbol{x}, 0) = \rho_0(\mathbf{x})$ and $\rho(\boldsymbol{x}, T) = \rho_T(\mathbf{x})$, then the classical L_2 Wasserstein distance can be shown to be equal to the infimum of:

$$\sqrt{\int_{\mathbf{R}^d} \int_0^T \rho(\boldsymbol{x}, t) |v(\boldsymbol{x}, t)|^2 \, d\boldsymbol{x} dt} \tag{1}$$

over all $v(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x}, t)$ which satisfy the continuity equation: $\frac{\partial \rho(\mathbf{x}, t)}{\partial t} = -\nabla \cdot (v(\mathbf{x}, t)\rho(\mathbf{x}, t))$. For the individual particles which make up this density field, this corresponds to a time-update in the position given by the vector field at their location, *i.e.*: $\frac{\partial \mathbf{x}}{\partial t} = v(\mathbf{x}, t)$. It turns out that, in terms of the velocity, the optimal solutions to eq. (1) can be written as the gradient of some potential function ϕ, *i.e.,* $v(\mathbf{x}, t) = \nabla_{\mathbf{x}} \phi(\mathbf{x}, t)$, thereby earning the name *potential flows*. Ultimately, by following such a potential flow, the system can be seen to be minimizing the Wasserstein distance, thereby solving the optimal transport problem.

In relation to latent traversals, we see that we can make an intuitive connection between the distribution of points which make up the start and end points of a given semantic traversal (*e.g.,* the distribution of portraits photos with short and long hair respectively), and the source and target distributions in the OT framework. Following such a connection would intuitively suggest that we may be able to learn a corresponding latent potential $\phi(\boldsymbol{x}, t)$ which defines the structure of the latent space with respect to this transformation, and then use the gradient of this field to move particles from one distribution to another.

While making a formal connection with OT remains beyond this paper, we see there is still a close intuitive connection between such methods which may be further formalized in future work. In this work, we present this connection simply as motivation for our method and empirically demonstrate the effectiveness and generality of our approach using this intuition. *One question which comes from this interpretation, is what kind of velocity fields are appropriate for encoding transformations?* In the following subsection, we provide further intuition that motivates our use of physically-realistic PDEs such as the wave equation to constrain the space-time dynamics of ϕ and the resulting velocity $\nabla \phi$.

2.2. Traveling Waves in Neuroscience

More abstractly, our work is motivated by the recent interest in traveling waves in the neuroscience literature. Succinctly, traveling waves have recently been observed to exist in a diversity of regions and scales in the biological cortex (Muller et al., 2018). Although a consensus has yet to be reached about their exact computational purpose, there is a variety of emerging work which appears to implicate them in the predictive processing of observed transformations from both biological (Jancke et al., 2004; Sato et al., 2012; Friston, 2019; Alamia & VanRullen, 2019; Besserve et al., 2015) and computational (Keller & Welling, 2023) perspectives. Specifically, these works suggest that they play the role of integrating information across time, encoding motion, and modulating information transfer. In this work, we leverage these observations to motivate the hypothesis that *traveling waves may be a neural correlate of latent traversals, and thereby serve as an efficient way to encode natural transformations using neural network architectures.* Pursuant to this hypothesis, we expect beneficial performance with physics-inspired PDEs guiding latent traversals in artificial neural networks as well.

3. Related Work

Latent Traversal in Generative Models. Latent traversals have often been used to evaluate the quality of learned latent spaces of the deep generative models (Kingma & Welling, 2014; Goodfellow et al., 2014). Pursuant to this, much research has been conducted to determine the optimal way to compute traversal trajectories in order to yield semantically meaningful generations. One line of research employs explicit human annotations to define the semantic labels for interpretable paths (Radford et al., 2015; Goetschalckx et al., 2019; Jahanian et al., 2020; Plumerault et al., 2020; Shen et al., 2020; Ling et al., 2021; Shi et al., 2022). By contrast, unsupervised methods discover interpretable directions without any prior knowledge (Härkönen et al., 2020; Kwon et al., 2023; Choi et al., 2022; Karmali et al., 2022; Spingarn-Eliezer et al., 2021; Ren et al., 2022; Oldfield et al., 2023). For example, Voynov & Babenko (2020) proposed to learn a set of semantic concepts via an auxiliary classifier. Other methods such as SeFa (Shen & Zhou, 2021) pointed out that the eigenvectors of the projection matrix following the latent codes can be directly used as interpretable directions. More recently, Tzelepis et al. (2021) proposed to non-linearly perturb the latent code using gradients of learned RBFs. Our work mainly belongs to the unsupervised category, as demonstrated by the majority of the results presented in Sec. 5; however, as we show in Sec. 4.3 and 5.4, our method can also be extended to the supervised setting, thereby regularizing the latent space towards increased structure and improving the model's ability to represent similarly structured transformations.

Disentanglement Learning. In contrast to the goal of discovering latent traversal trajectories in pre-trained models, other methods have aimed to attain an a priori structured representation through additional regularization during training. For example, InfoGAN (Chen et al., 2016) encouraged disentanglement by maximizing the mutual information between the observations and a fixed subset of the latent code. Zhu et al. (2020) proposed a variational predictability loss to learn disentangled representations and introduced a metric to evaluate unsupervised disentanglement methods. Peebles et al. (2020); Wei et al. (2021) and Song et al. (2022) proposed different orthogonality constraints to improve disentanglement ability. Alternatively, for disentanglement with VAEs, much work has focused on various modifications to the evidence lower bound (ELBO) to encourage increased independence of the different latent dimensions. Most notably, the β-VAE (Higgins et al., 2016) first introduced a hyper-parameter to accentuate the penalty of the divergence between the prior and variational posterior. Follow-up research used additional guidance to encourage improved disentanglement in this manner, including β-TC-VAE (Kim & Mnih, 2018; Chen et al., 2018a), DIP-VAE (Kumar et al., 2018), Guided-VAE (Ding et al., 2020), JointVAE (Dupont, 2018), and CasadedVAE (Jeong & Song, 2019).

Physics for Deep Learning. In recent years, an increased effort has developed to combine deep neural networks with concepts from physics. Much work has focused on using deep learning to solve problems that arise in physics, such as solving PDEs by Physics Informed Neural Networks (PINNs) (Raissi et al., 2019), learning dynamic systems with Neural ODEs (Chen et al., 2018b), and discovering physical concepts (Iten et al., 2020). Another active research field leverages fundamental laws (*e.g.,* symmetries or conservation laws) to improve deep learning models. Some examples include designing equivariant neural networks to handle input with geometric symmetries (Cohen & Welling, 2016; Cohen et al., 2018; Zhang, 2019; Satorras et al., 2021; Keller & Welling, 2021), endowing neural networks with Hamiltonian dynamics for improved performance and generalization (Greydanus et al., 2019; Toth et al., 2020), and building score-based denoising diffusion models for generative modelling (Ho et al., 2020; Song et al., 2021a;b). In this work, we use PINN-inspired constraints to model the latent traversal with learned potential PDEs, situating our model in the category of work which seeks to improve deep learning with physically inspired methods.

4. Methodology

In this section we present the formulation of our learned potential functions, their integration into generative models under different settings, and the training and sampling strategies. The overview of our method is depicted in Fig. 1.

4.1. Latent Traversals as Potential Flows

Learning the Potential PDE. Assume we are given a pre-trained generative model $\mathcal{G} : \mathcal{Z} \rightarrow \mathcal{X}$ with prior distribution

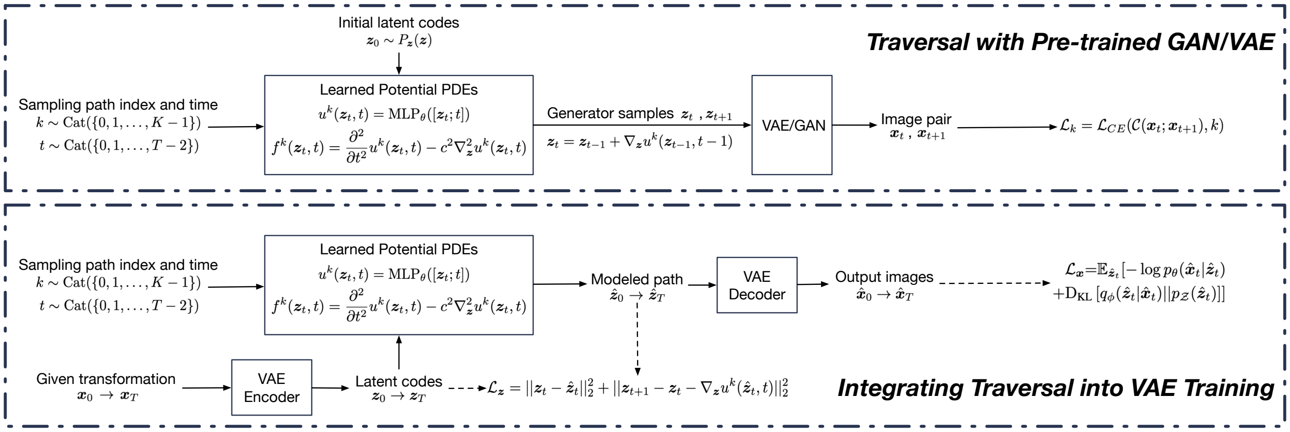
Initial latent codes
$z_0 \sim P_z(z)$
Traversal with Pre-trained GAN/VAE
Sampling path index and time
$k \sim \mathrm{Cat}(\{0, 1, \ldots, K-1\})$
$t \sim \mathrm{Cat}(\{0, 1, \ldots, T-2\})$
Learned Potential PDEs
$u^k(z_t, t) = \mathrm{MLP}_\theta([z_t; t])$
$f^k(z_t, t) = \frac{\partial^2}{\partial t^2} u^k(z_t, t) - c^2 \nabla_z^2 u^k(z_t, t)$
Generator samples z_t , z_{t+1}
$z_t = z_{t-1} + \nabla_z u^k(z_{t-1}, t-1)$
VAE/GAN
Image pair
x_t , x_{t+1}
$\mathcal{L}_k = \mathcal{L}_{CE}(\mathcal{C}(x_t; x_{t+1}), k)$
Sampling path index and time
$k \sim \mathrm{Cat}(\{0, 1, \ldots, K-1\})$
$t \sim \mathrm{Cat}(\{0, 1, \ldots, T-2\})$
Learned Potential PDEs
$u^k(z_t, t) = \mathrm{MLP}_\theta([z_t; t])$
$f^k(z_t, t) = \frac{\partial^2}{\partial t^2} u^k(z_t, t) - c^2 \nabla_z^2 u^k(z_t, t)$
Modeled path
$\hat{z}_0 \to \hat{z}_T$
VAE
Decoder
Output images
$\hat{x}_0 \to \hat{x}_T$
$\mathcal{L}_x = \mathbb{E}_{\hat{z}_t}[-\log p_\theta(\hat{x}_t|\hat{z}_t)$
$+ \mathrm{D}_{\mathrm{KL}}[q_\phi(\hat{z}_t|\hat{x}_t)||p_Z(\hat{z}_t)]]$
Given transformation
$x_0 \to x_T$
VAE
Encoder
Latent codes
$z_0 \to z_T$
$\mathcal{L}_z = ||z_t - \hat{z}_t||_2^2 + ||z_{t+1} - z_t - \nabla_z u^k(\hat{z}_t, t)||_2^2$
Integrating Traversal into VAE Training

Figure 1. Overview of our learned potential PDEs for latent traversal in two different experimental settings.

$P_z(z)$. To model K different semantically disentangled latent trajectories, we model each trajectory separately as the gradient of a learned time-dependant scalar potential energy field: $u^k(z_t, t) = \mathrm{MLP}_{\theta^k}([z_t; t]) \in \mathbb{R}$. In this work we use a small multilayer perceptron (MLPs) to learn each potential. The process of traversing from an initial sample (z_0) to a future element (z_t) at time t is then defined as the potential flow $\nabla_z u$ described by this field:

$$z_0 \sim P_z(z) \qquad z_t = z_{t-1} + \nabla_z u^k(z_{t-1}, t-1) \tag{2}$$

To encourage the latent potential to model realistic trajectories and follow the intuitions outlined above, we additionally impose a PINN constraint in the form of the second-order wave equation with wave coefficient c:

$$f^k(z_t, t) = \frac{\partial^2}{\partial t^2} u^k(z_t, t) - c^2 \nabla_z^2 u^k(z_t, t) \tag{3}$$

Such a constraint makes our potential flow model a good approximation of small amplitude sound waves (Lamb, 1993), and empirically is seen to produce highly diverse and realistic trajectories. Our objective is then to minimize:

$$\mathcal{L}_f = \frac{1}{T} \sum_{t=0}^{T-1} ||f^k(z_t, t)||_2^2, \; \mathcal{L}_u = ||\nabla_z u^k(z_0, 0)||_2^2 \tag{4}$$

where T represents the total number of timesteps of our latent trajectory, $\mathcal{L}_f$ restricts the energy to obey our physical constraints, and $\mathcal{L}_u$ restricts $u(z_t, t)$ to return no update at t=0, thereby matching the initial condition.

Jacobian Regularization. While the above formulation models traversals as physically realistic potential flows, it cannot ensure that the modeled traversal paths are semantically meaningful. Therefore, to make our learned potentials more aligned with the semantics of the data, we take inspiration from prior work and further couple the traversal direction with the Jacobian of the generator. Similar to Zhu et al. (2021; 2022), we first approximate the manipulation on the latent space as

$$\mathcal{G}(z_t + \epsilon \nabla u^k(z_t, t)) \approx \mathcal{G}(z_t) + \epsilon \underline{\frac{\partial \mathcal{G}(z_t)}{\partial z_t} \nabla_z u^k(z_t, t)} \tag{5}$$

where ϵ denotes perturbation strength. Intuitively, for sufficiently small ϵ, if the Jacobian-vector product (the underlined term in eq. (5)) can cause large variations in the generated sample, the direction is likely to be semantically meaningful. We therefore introduce a Jacobian-vector product regularization term to encourage the improved semantic variations of our traversals in an unsupervised manner:

$$\mathcal{L}_\mathcal{J} = -||\frac{\partial \mathcal{G}(z_t)}{\partial z_t} \nabla_z u^k(z_t, t)||_2^2 \tag{6}$$

4.2. Traversal with Pre-trained GAN/VAE

With pre-trained models, the weights of the generator are frozen. We only update the parameters of our MLPs and of the auxiliary potential-index classifier module. We adopt an auxiliary classifier $\mathcal{C}$ to predict the potential index and use the cross-entropy loss to optimize it:

$$\hat{k} = \mathcal{C}(x_t; x_{t+1}), \; \mathcal{L}_k = \mathcal{L}_{CE}(\hat{k}, k) \tag{7}$$

Where $x_t = \mathcal{G}(z_t)$ is the generated sample from timestep t.

4.3. Integrating Traversal into VAE Training

When training VAEs from scratch, our method can perform "supervised" latent traversal as extra regularization to improve the likelihood. That is, we explicitly model the path of the variations of a semantic attribute during the training process. In this setting, we consider having access to the pre-defined transformation of each variation factor $x_0 \to x_T$. Then we can obtain the corresponding latent codes $z_0 \to z_T$ by feeding images to the encoder, *i.e.,*

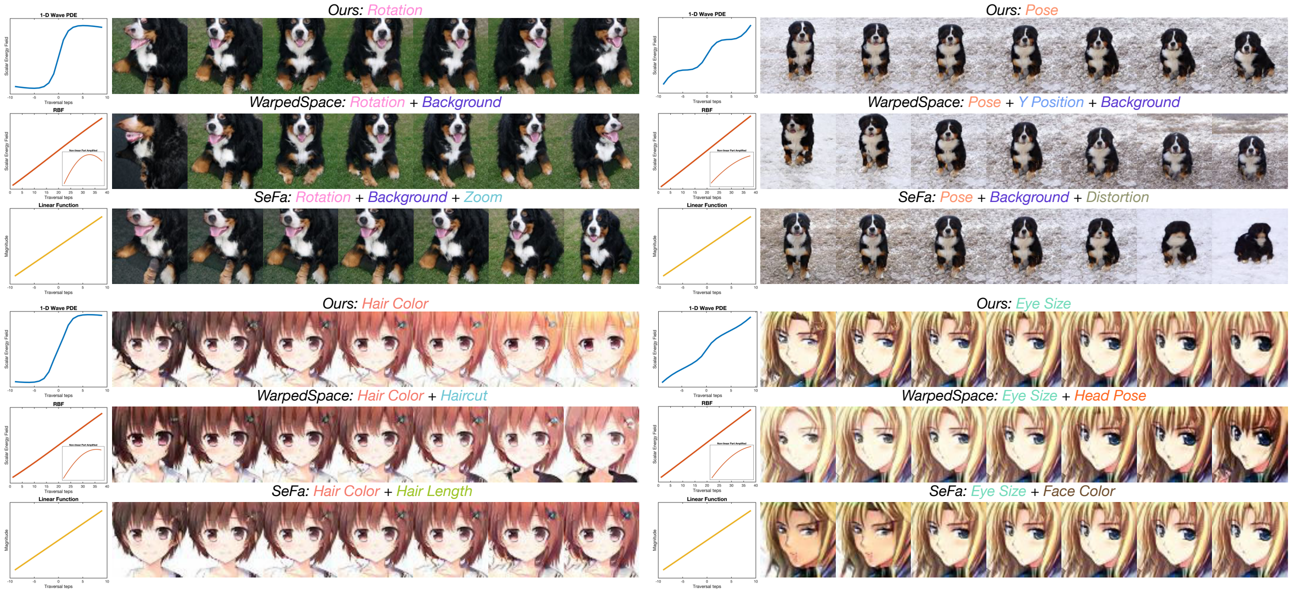

Figure 2. Exemplary traversal paths (potential PDEs for our method) and the corresponding interpolation images with SNGAN and BigGAN. Since the paths of WarpedSpace are of very limited non-linearity that is hard to perceive, we amplify the non-linear part in the sub-figure inside the figure as follows: for a traversal path $\boldsymbol{y}$ of WarpedSpace, we decompose it into $\boldsymbol{y} = \boldsymbol{y}_{LN} + \boldsymbol{y}_{NLN}$ where $\boldsymbol{y}_{LN}$ denotes the linear part and $\boldsymbol{y}_{NLN}$ is the non-linear counterpart. Then the non-linearity part is amplified by $\boldsymbol{y} = \boldsymbol{y}_{LN} + 200 \cdot \boldsymbol{y}_{NLN}$.

$\boldsymbol{z}_t = \texttt{Encode}(\boldsymbol{x}_t)$. Then our potential PDEs manipulate the initial latent codes $\boldsymbol{z}_0$ to obtain $\hat{\boldsymbol{z}}_1 \rightarrow \hat{\boldsymbol{z}}_T$ by progressively performing $\hat{\boldsymbol{z}}_t = \boldsymbol{z}_0 + \sum \nabla_{\boldsymbol{z}} u^k$. The output images $\hat{\boldsymbol{x}}_1 \rightarrow \hat{\boldsymbol{x}}_T$ can be easily attained by decoding $\hat{\boldsymbol{z}}_1 \rightarrow \hat{\boldsymbol{z}}_T$. The traversal paths modeled by our wave equations are encouraged to match the ground truth as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{z}} &= ||\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t||_2^2 + ||(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t) - (\hat{\boldsymbol{z}}_{t+1} - \hat{\boldsymbol{z}}_t)||_2^2 \\ &= ||\boldsymbol{z}_t - \hat{\boldsymbol{z}}_t||_2^2 + ||\boldsymbol{z}_{t+1} - \boldsymbol{z}_t - \nabla_{\boldsymbol{z}} u^k(\hat{\boldsymbol{z}}_t, t)||_2^2\end{aligned} \tag{8}$$

where the first term penalizes the difference between current latent codes and the ground truth history, and the second term ensures that the future update at the next timestep is realistic. Besides improving the plausibility of traversal paths, we optimize the ELBO:

$$\mathcal{L}_{\boldsymbol{x}} = \mathbb{E}_{\hat{\boldsymbol{z}}_t}[-\log p_\theta(\hat{\boldsymbol{x}}_t|\hat{\boldsymbol{z}}_t) + \mathrm{D}_{\mathrm{KL}}[q_\phi(\hat{\boldsymbol{z}}_t|\hat{\boldsymbol{x}}_t)||p_{\mathcal{Z}}(\hat{\boldsymbol{z}}_t)]] \tag{9}$$

where p_θ parameterizes the generator, and q_ϕ denotes the approximate posterior. The combination of the two losses could yield more structured latent space and more realistic traversal trajectories, which might improve the likelihood.

4.4. Sampling and Training Strategies

At each training step, we randomly sample a potential index k from $\mathrm{Cat}(\{0, 1, \dots, K{-}1\})$ and a timestep t from $\mathrm{Cat}(\{0, 1, \dots, T{-}2\})$. Then we use the selected potential to generate the corresponding velocity fields and obtain the two latent codes $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t+1}$. Subsequently, the generator is fed with the latent codes and outputs a pair of images $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$. Finally, we adopt an auxiliary classifier to predict the potential index $\hat{k}$. The overall loss function is defined as

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_f + \underline{\mathcal{L}_{\mathcal{J}} + \mathcal{L}_k} + \boxed{\mathcal{L}_{\boldsymbol{x}} + \mathcal{L}_{\boldsymbol{z}}} \tag{10}$$

where $\mathcal{L}_k$ matches the predicted index $\hat{k}$ to the ground truth k, therefore encouraging that each learned potential is significantly distinct and self-consistent to be recognized by a classifier accurately. The boxed terms are only applied to regularize the latent space when integrated into VAE training, while the underlined terms are used for pre-trained models. Notice that different from Voynov & Babenko (2020); Tzelepis et al. (2021), we do not predict the timesteps from the image pair $[\boldsymbol{x}_t, \boldsymbol{x}_{t+1}]$. This is because our potential PDEs can be very diverse in spatiotemporal form, thus predicting the timesteps from two points on the path demonstrated to be both unnecessary and practically infeasible.

5. Experiments

This section starts with the setup, followed by the results under different settings, and ends with in-depth discussions.

5.1. Settings

Models and Datasets. For experiments of pre-trained GANs, our method is evaluated on SNGAN (Miyato et al., 2018) with AnimeFace (Chao, 2019), BigGAN (Brock et al., 2019) with ImageNet (Deng et al., 2009), and Style-

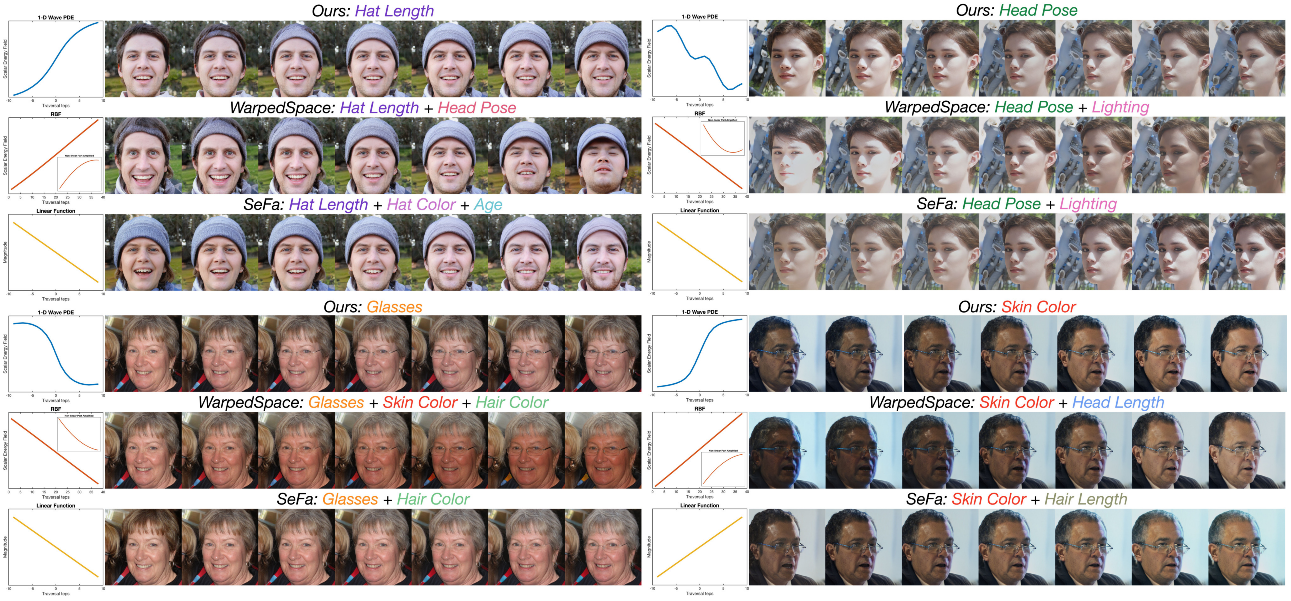

Figure 3. Traversal trajectories (potential PDEs for our method) and the associated interpolation images of the exemplary four attributes with StyleGAN2. The non-linearity of WarpedSpace paths is amplified in the same way as done in SNGAN and BigGAN.

GAN2 (Karras et al., 2020) with FFHQ (Karras et al., 2019). For BigGAN, we train the target class "Bernese mountain dog". We adopt LeNet (LeCun et al., 1998) as the auxiliary classifier for SNGAN, while ResNet-18 (He et al., 2016) based classifier is used for both BigGAN and StyleGAN2. For the VAEs experiments, we use the VAE encoder as the auxiliary classifier and evaluate our method on MNIST (LeCun, 1998) and dSprites (Matthey et al., 2017) datasets.

MLP for Modeling PDEs. We use sinusoidal positional embeddings (Vaswani et al., 2017) to embed the timestep t. Linear layers with `Tanh` activations are used for embedding the latent code input $\boldsymbol{z}$. Another linear layer is used to fuse features across space and time. We set the wave coefficient c as a learnable parameter and initialize it with 1.

Metrics. For the quantitative evaluation of traversal with GANs, we use Variational Predictability (VP) (Zhu et al., 2020) score and the correlation coefficient between face attributes and traversal steps using pre-trained attribute estimators. The VP score adopts the few-shot learning setting (*e.g.,* 10% images as the training set) to measure the generalization of a simple neural network in classifying the discovered latent directions from a crafted dataset of random image pairs $[\boldsymbol{x}_0, \boldsymbol{x}_T]$. For attribute correlation, we first use S3FD (Zhang et al., 2017) to extract the face region and then compute the normalized Pearson's correlation between potential indexes and traversal steps using several pre-trained attributes estimators, including ArcFace (Deng et al., 2019) for face identity, FairFace (Karkkainen & Joo, 2021) for face attributes (age, race, and gender), and HopeNet (Doosti et al., 2020) for face poses (yaw, pitch, and roll). The correlation results are averaged across 50 random latent samples. For the quantitative evaluation of VAEs, since our method performs vector-based manipulation, traditional single-dimension-based VAE disentanglement metrics such as Mutual Information Gap (MIP) (Chen et al., 2018a) do not apply here. Some works such as Arvanitidis et al. (2018); Tonnaer et al. (2020) can perform the evaluation of quantitative vector-based manipulation but they require supervision of the ground truth. We thus also evaluate the disentanglement performance using the VP score. The log-likelihood over the entire dataset is measured for the experiment of integrating our method into the VAE training.

Baselines. For pre-trained GANs, we compare our method against two representative baselines, *i.e.,* SeFa (Shen & Zhou, 2021) and WarpedSpace (Tzelepis et al., 2021). SeFa uses eigenvectors of the weight matrix after latent codes for *linear* perturbation, while WarpedSpace *non-linearly* changes the latent codes using the gradients of RBFs. As for VAEs, there are no popular vector-based traversal methods in the literature so we also use WarpedSpace for comparison. Finally, as another controlled baseline, we train a linear function with other settings aligned with our method.

Table 1. Comparison of the VP scores (%) with different GANs. The results are averaged over 3 random runs.

Models	SeFa	WarpedSpace	Ours
SNGAN	53.76	58.83	**65.89**
BigGAN	13.59	14.07	**15.29**
StyleGAN2	39.20	36.31	**48.54**

Table 2. The l_1 normalized attribute correlations of our method (*top*), WarpedSpace (*middle*), and SeFa (*bottom*) based on 50 samples. The second highest correlation is also highlighted if the best value in the row is not on the diagonal.

	Yaw	Pitch	Roll	Identity	Age	Race	Gender
Yaw	**0.34**	0.09	0.22	0.09	0.03	0.18	0.03
Pitch	0.04	**0.25**	0.11	0.08	0.00	0.08	**0.45**
Roll	0.23	0.19	**0.35**	0.00	0.02	0.03	0.18
Identity	0.01	0.06	0.00	**0.61**	0.21	0.03	0.07
Age	0.00	0.06	0.00	0.03	**0.87**	0.00	0.04
Race	0.05	0.07	0.06	0.02	0.01	**0.73**	0.06
Gender	0.08	0.19	0.09	0.04	0.00	0.03	**0.58**

	Yaw	Pitch	Roll	Identity	Age	Race	Gender
Yaw	**0.34**	0.03	0.05	**0.42**	0.01	0.08	0.07
Pitch	0.01	**0.38**	0.07	**0.42**	0.01	0.09	0.01
Roll	0.10	0.15	0.17	**0.27**	0.02	0.07	**0.22**
Identity	0.01	0.10	0.00	**0.69**	0.10	0.07	0.01
Age	0.02	0.09	0.05	**0.52**	**0.25**	0.02	0.05
Race	0.05	0.02	0.07	0.12	0.07	**0.54**	0.12
Gender	0.09	0.00	0.02	0.40	0.00	0.00	**0.49**

	Yaw	Pitch	Roll	Identity	Age	Race	Gender
Yaw	**0.29**	0.01	0.05	**0.40**	0.04	0.09	0.11
Pitch	0.09	**0.29**	0.06	**0.41**	0.05	0.08	0.01
Roll	0.03	0.10	0.09	**0.60**	0.00	0.06	**0.12**
Identity	0.02	0.05	0.02	**0.74**	0.08	0.08	0.01
Age	0.02	0.08	0.02	**0.47**	**0.25**	0.02	0.15
Race	0.07	**0.25**	0.02	**0.58**	0.00	0.00	0.07
Gender	0.02	0.05	0.02	**0.43**	0.02	**0.35**	0.12

5.2. Results with Pre-trained GANs

SNGAN and BigGAN. Fig. 2 displays the exemplary latent traversal results and the corresponding trajectories with SNGAN and BigGAN. Since the parameters of the generator are frozen, each method would generate the same image for one latent sample. Our PDEs can generate traversal paths with distinct semantics and precise image attribute control, while the baselines suffer from entangled attributes and the non-target semantics also vary during traversal. Moreover, the paths of WarpedSpace are of very limited non-linearity, which is imperceptible unless the non-linear part of the path is significantly amplified. By contrast, our potential PDEs have more diverse shapes and more flexible non-linearity. Table 1 presents the quantitative evaluation results of the VP scores. Our PDEs achieve state-of-the-art performance in terms of classification accuracy in the few-shot learning setting. Specifically, our method outperforms the second-best baseline by 7.04% with SNGAN, by 1.22% with BigGAN, and by 12.23% with StyleGAN2. The consistent performance gain on each dataset indicates that the semantics of our traversal paths are indeed more disentangled than others. It is also worth mentioning that the relatively marginal advantage with BigGAN might stem from the fact that BigGAN generates images in wide domains ($1,000$ ImageNet classes). This domain diversity might restrict the actual number of latent semantics, thus limiting the performance.

StyleGAN2. Fig. 3 compares the exemplary latent traversal with StyleGAN2. The results are coherent with those on SNGAN and BigGAN: the traversal paths of baselines suffer from entangled semantics, while our potential PDEs are able to model trajectories that correspond to more disentangled image attributes. Table 2 presents the l_1 normalized correlation results of some common face attributes. As can be seen, most attributes of both SeFa and WarpedSpace have the highest correlation with "identity", implying that their variations of these attributes are often coupled with variations of the face identity during the traversal. By contrast, our method has the best attribute correlations mostly on the diagonal, which explicitly indicates that these attributes of our method are more disentangled from each other.

Table 3. Comparison of the VP scores (%) with pre-trained VAEs. The results are averaged over 3 random runs.

Models	**WarpedSpace**	**Ours (Linear)**	**Ours**
MNIST	13.44	12.76	**17.38**
dSprites	15.01	14.25	**18.49**

5.3. Results with Pre-trained VAEs

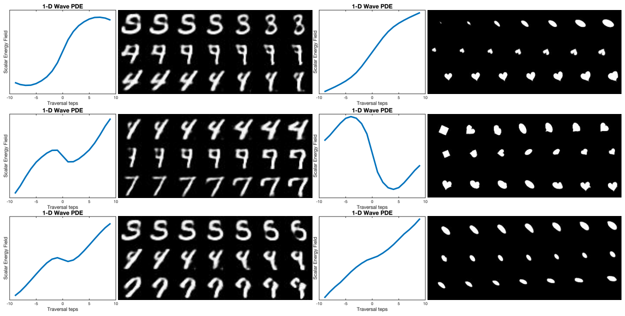

Figure 4. Exemplary semantic attributes and the corresponding traversal trajectories with VAEs trained on MNIST and dSprites.

Fig. 4 displays the exemplary semantics discovered by our method with pre-trained VAEs. Our potential PDEs exhibit a diverse set of different shapes and the interpolation images correspond to distinct transformation factors. Table 3 presents the quantitative evaluation of VP scores. The linear baseline and WarpedSpace achieve similar performance, falling behind our method by 4%. This demonstrates again the effectiveness of our PDEs in modelling latent traversal.

Table 4. The log-likelihood $\log p_\theta(\boldsymbol{x})$ evaluated over the dataset.

Models	**Naively Trained**	**Trained with Our Method**
MNIST	-2207.70	**-2144.71**
dSprites	-3848.04	**-3740.97**

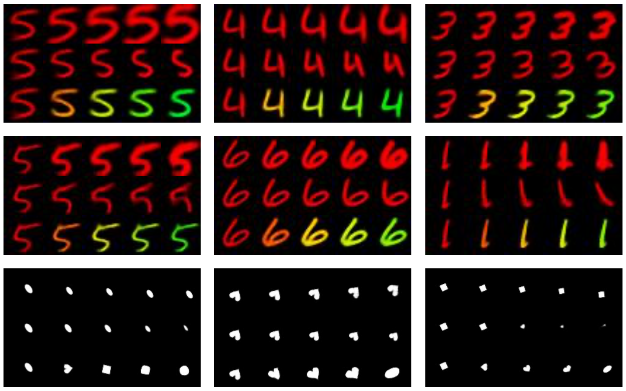

Figure 5. Exemplary traversal results when our method is integrated into the VAE training process. For MNIST, the exhibited transformations are scaling, rotation, and coloring changes from top to bottom. For Dsrpites, the corresponding transformations are y-axis position, scaling, and shape changes from top to bottom.

5.4. Results with VAEs Trained from Scratch

Table 4 compares the log-likelihood of VAEs integrated with our method. Notice that common disentanglement methods would often sacrifice the likelihood (Higgins et al., 2016). However, integrating our PDEs into the training process slightly improves the likelihood estimation. Fig. 5 displays the exemplary traversal results of the pre-defined transformations. Our method is also able to learn and generalize the pre-defined transformation factors well.

Table 5. Equivariance error on MNIST.

Transformations	Rotation	Scaling	Coloring
Our Method	**235.96**	**230.39**	**240.64**
Vanilla VAE	1278.21	1309.56	1370.54

One interesting geometric property induced by our potential flows is the approximate equivariance for VAEs trained from scratch. At a high level, an equivariant map is one which commutes with a desired transformation group, *i.e.,* $T'[f(x)] = f(T[x])$. This can be understood as preserving geometric symmetries of the input space. The gradient of our potential function can be interpreted as the equivariant latent operator T' corresponding to the observed input transformation $T[x]$. As is typical in the equivariance literature, we can measure how close this is to exact equivariance by measuring the equivariance error:

$$\begin{aligned}\texttt{Err} &= \sum_{t=1}^{T} |\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t| \\ &= \sum_{t=1}^{T} |\boldsymbol{x}_t - \texttt{Decode}(\boldsymbol{z}_0 + \sum^{t} \nabla_{\boldsymbol{z}} u^k)|\end{aligned} \tag{11}$$

We see this is equivalent to measuring the satisfaction of the equivariance relation $T[x] - f^{-1}(T'[f(x)]) = 0$ where f^{-1} is approximated with the decoder. Table 5 presents the evaluation results against a vanilla VAE on transforming MNIST. Note that since the vanilla VAE has no notion of a corresponding transformation in the latent space T' (*i.e.,* no a priori known latent structure), we simply set $\nabla_{\boldsymbol{z}} u^k$ to 0 and treat this as a lower bound baseline. We see that our method performs significantly above this baseline, indicating that it could be helpful to build equivariant VAEs.

5.5. Discussions

Linear Directions as Special Cases. We note that the linear traversal approaches can be understood as special cases of our second-order wave equations. Actually, for general linear functions defined as $u(x, t) = a \cdot x + b \cdot t$ where a and b denote the coefficients, the solutions would all correspond to wave equations. In this sense, linear functions are simplified special cases of our waves. One piece of evidence for supporting this is that in certain cases where the structure of the latent space might be simple, our PDEs can also reduce back to functions that are almost linear, such as the traversal paths of the semantic attribute "Eye Size" in Fig. 2 and the transformation of scaling in Fig. 4 right.

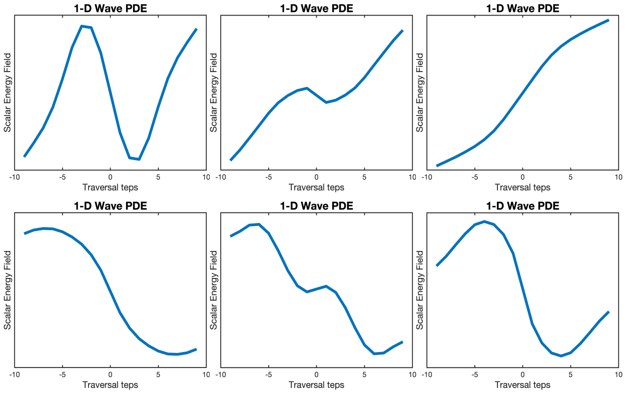

Figure 6. Common shapes of potential PDEs in our experiments.

Path Diversity. Our potential PDEs can be very different in shape and period. Fig. 6 exhibits some common PDEs learned in our experiments. As can be seen, our wave equations allow for a wide set of traversal paths, ranging from linear lines to traveling waves of a full period. This flexibility enables modeling diverse trajectories in the manifold.

Semantic and Trajectory Unambiguity. As shown in Fig. 7, for the same traversal path, the semantic attribute is consistent to different samples and the corresponding PDE paths are of very similar shapes. Take the semantic attribute of "Zoom IN" as an example. The scalar potential energy fields of the three images all have slow changes near the endpoints while taking sharp increases in the middle regime. Accordingly, the interpolation images coincide with

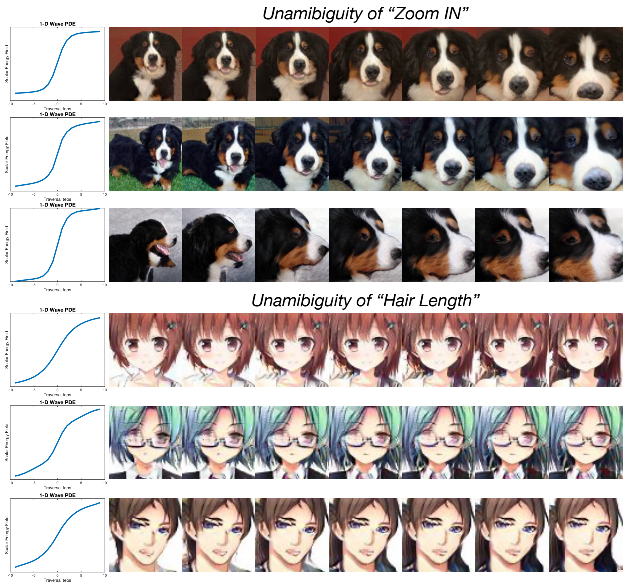

Figure 7. Unambiguity of our potential PDEs and the corresponding discovered semantics: the shape of trajectory and the image attribute of a traversal path are consistent to different samples.

identical semantics.

Geometric Properties of Latent Spaces. Besides the equivariance property of the encoder/decoder, we also have some novel observations about the shape and variations of $\nabla_{\boldsymbol{z}} u^k$. For VAEs, we observe that the ***simple*** variation factors that involve ***linear*** transformations (*e.g.*, scaling and translation shown in Fig. 4 right) tend to be accordingly ***more linear*** in the latent space. For GANs, the semantic attributes that edit ***local*** image regions tend to be ***more linear*** in the latent space, such as the attribute "Eye Size" in Fig. 2 and the attributes "Glasses" and "Hat Length" in Fig. 3. Through all the experiments, the traversal directions generally tend to have fewer variations when closer to the endpoints. We think this is because at the endpoints (*i.e.*, large timesteps) our potentials learnt to not violate the semantic attribute and not to go out of the data manifold.

Limitations of Potential Flows. It is known that potential flows are limited in their ability to represent all forms of physically known flows. For example, since the curl of the gradient is known to be zero, potential flows are inherently irrotational and thus cannot model vorticity. In the case of latent traversals, the literature largely appears to model non-cyclic transformations (such as hair length or skin color), and thus this modeling assumption is observed to be valid. However, this limitation explains why the rotation traversals attempted to be learned by our VAE model perform poorly. Ultimately, we propose this framework as a first step towards modeling latent traversals with more complex, physically informed dynamics, and suggest that in some settings, these physical biases may match the underlying data in a beneficial way. We propose that valuable future work could explore alternative parameterizations of the latent vector field which could respectively yield alternative biases suitable to other datasets.

Alternative PDE Modeling Approaches. We mainly explore the PINN-based physical constraints to model our PDEs. Despite the flexibility and efficiency, this approach achieves the *soft* PDE constraints approximately. Other alternative possibilities for PDE modeling include Neural Conservation Laws (Richter-Powell et al., 2022) that impose *hard* divergence-free constraints and accurate neural PDE solvers (Hsieh et al., 2019; Brandstetter et al., 2022). Investigating other PDE modeling approaches is an important research direction in future work.

Famous PDEs of the Sample Evolution. Driven by our learned velocity field $\nabla u(\boldsymbol{z}, t)$, the sample evolution of $\boldsymbol{z}$ over space and time could satisfy certain PDEs. In particular, with certain $\nabla u(\boldsymbol{z}, t)$, the evolution of $\boldsymbol{z}$ could possibly become some special well-known PDEs, such as heat equations, Fokker Planck equations, and Porous Medium equations. The specific types depend on the relation between $\nabla u(\boldsymbol{z}, t)$ and $\rho(\boldsymbol{z}, t)$. For instance, if the velocity field is set as $\nabla u(\boldsymbol{z}, t) = -\nabla \log(\rho(\boldsymbol{z}, t))$, the evolution would become the heat equations. More details about the possible relations are kindly referred to Santambrogio (2017).

6. Conclusion

Inspired by the fluid mechanical interpretation of optimal transport and the role of traveling waves in neuroscience, we propose to model the latent traversal flexibly by the gradient flows of learned dynamic potential landscapes. Our method can model a set of traversal paths with distinct semantics to improve the disentanglement ability of pre-trained GANs and VAEs. Furthermore, our PDEs can be integrated into the training process of VAEs as regularization on the latent space to improve the model likelihood estimation.

Acknowledgements

This research was supported by the EU H2020 project AI4Media (No. 951911). Yue Song acknowledges travel support from ELISE (GA no 951847). Andy Keller thanks the Bosch Center for Artificial Intelligence for funding.

References

Alamia, A. and VanRullen, R. Alpha oscillations and traveling waves: Signatures of predictive coding? *PLoS Biology*, 2019. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000487.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. *ICLR*, 2018. URL https://arxiv.org/abs/1710.11379.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. URL https://link.springer.com/article/10.1007/s002110050002.

Besserve, M., Lowe, S. C., Logothetis, N. K., Schölkopf, B., and Panzeri, S. Shifts of gamma phase across primary visual cortical sites reflect dynamic stimulus-modulated information transfer. *PLoS biology*, 2015. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002257.

Brandstetter, J., Worrall, D., and Welling, M. Message passing neural pde solvers. *ICLR*, 2022. URL https://arxiv.org/abs/2202.03376.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. URL https://arxiv.org/abs/1809.11096.

Chao, B. Anime face dataset: a collection of high-quality anime faces., 2019. URL https://github.com/bchao1/Anime-Face-Dataset.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. *NeurIPS*, 2018a. URL https://arxiv.org/abs/1802.04942.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *NeurIPS*, 2018b. URL https://arxiv.org/abs/1806.07366.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 2016. URL https://arxiv.org/abs/1606.03657.

Choi, J., Lee, J., Yoon, C., Park, J. H., Hwang, G., and Kang, M. Do not escape from the manifold: Discovering the local coordinates on the latent space of gans. *ICLR*, 2022. URL https://arxiv.org/abs/2106.06959.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *ICML*, 2016. URL https://arxiv.org/abs/1602.07576.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. *ICLR*, 2018. URL https://arxiv.org/abs/1801.10130.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. URL https://ieeexplore.ieee.org/document/5206848.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. URL https://arxiv.org/abs/1801.07698.

Ding, Z., Xu, Y., Xu, W., Parmar, G., Yang, Y., Welling, M., and Tu, Z. Guided variational autoencoder for disentanglement learning. In *CVPR*, 2020. URL https://arxiv.org/abs/2004.01255.

Doosti, B., Naha, S., Mirbagheri, M., and Crandall, D. J. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, 2020. URL https://arxiv.org/abs/2004.00060.

Dupont, E. Learning disentangled joint continuous and discrete representations. *NeurIPS*, 2018. URL https://arxiv.org/abs/1804.00104.

Friston, K. J. Waves of prediction. *PLoS biology*, 2019. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000426.

Goetschalckx, L., Andonian, A., Oliva, A., and Isola, P. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. URL https://arxiv.org/abs/1906.10112.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, 2014. URL https://arxiv.org/abs/1406.2661.

Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. *NeurIPS*, 2019. URL https://arxiv.org/abs/1906.01563.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. *NeurIPS*, 2020. URL http://128.84.4.34/abs/2004.02546.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016. URL https://arxiv.org/abs/1512.03385.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2016. URL https://openreview.net/forum?id=Sy2fzU9gl.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.11239.

Hsieh, J.-T., Zhao, S., Eismann, S., Mirabella, L., and Ermon, S. Learning neural pde solvers with convergence guarantees. *ICLR*, 2019. URL https://arxiv.org/abs/1906.01200.

Iten, R., Metger, T., Wilming, H., Del Rio, L., and Renner, R. Discovering physical concepts with neural networks. *Physical review letters*, 124(1):010508, 2020. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.124.010508.

Jahanian, A., Chai, L., and Isola, P. On the" steerability" of generative adversarial networks. *ICLR*, 2020. URL https://arxiv.org/abs/1907.07171.

Jancke, D., Chavane, F., Naaman, S., and Grinvald, A. Imaging cortical correlates of illusion in early visual cortex. *Nature*, 2004. URL https://www.nature.com/articles/nature02396.

Jeong, Y. and Song, H. O. Learning discrete and continuous factors of data via alternating disentanglement. In *ICML*, 2019. URL https://arxiv.org/abs/1905.09432.

Karkkainen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021. URL https://arxiv.org/abs/1908.04913.

Karmali, T., Parihar, R., Agrawal, S., Rangwani, H., Jampani, V., Singh, M., and Babu, R. V. Hierarchical semantic regularization of latent spaces in stylegans. In *ECCV*, 2022. URL https://arxiv.org/abs/2208.03764.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. URL https://arxiv.org/abs/1812.04948.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. URL https://arxiv.org/abs/1912.04958.

Keller, T. A. and Welling, M. Topographic vaes learn equivariant capsules. *NeurIPS*, 2021. URL https://arxiv.org/abs/2109.01394.

Keller, T. A. and Welling, M. Locally coupled oscillatory recurrent neural networks learn traveling waves and topographic organization. *Cosyne abstracts*, 2023.

Kim, H. and Mnih, A. Disentangling by factorising. In *ICML*, 2018. URL https://arxiv.org/abs/1802.05983.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *ICLR*, 2014. URL https://arxiv.org/abs/1312.6114.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *ICLR*, 2018. URL https://arxiv.org/abs/1711.00848.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. *ICLR*, 2023. URL https://arxiv.org/abs/2210.10960.

Lamb, H. *Cambridge mathematical library: Hydrodynamics*. Cambridge University Press, Cambridge, England, 6 edition, November 1993.

LeCun, Y. The mnist database of handwritten digits. 1998. URL http://yann.lecun.com/exdb/mnist/.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. URL https://ieeexplore.ieee.org/abstract/document/726791.

Ling, H., Kreis, K., Li, D., Kim, S. W., Torralba, A., and Fidler, S. Editgan: High-precision semantic image editing. *NeurIPS*, 2021. URL https://arxiv.org/abs/2111.03186.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset, 2017. URL https://github.com/deepmind/dsprites-dataset/.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *ICLR*, 2018. URL https://arxiv.org/abs/1802.05957.

Muller, L., Chavane, F., Reynolds, J., and Sejnowski, T. J. Cortical travelling waves: mechanisms and computational principles. *Nature Reviews Neuroscience*, 2018. URL https://www.nature.com/articles/nrn.2018.20.

Oldfield, J., Tzelepis, C., Panagakis, Y., Nicolaou, M. A., and Patras, I. Panda: Unsupervised learning of parts and appearances in the feature maps of gans. *ICLR*, 2023. URL https://arxiv.org/abs/2206.00048.

Peebles, W., Peebles, J., Zhu, J.-Y., Efros, A., and Torralba, A. The hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*, 2020. URL https://arxiv.org/abs/2008.10599.

Plumerault, A., Borgne, H. L., and Hudelot, C. Controlling generative models with continuous factors of variations. *ICLR*, 2020. URL https://arxiv.org/abs/2001.10238.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2015. URL https://arxiv.org/abs/1511.06434.

Raissi, M., Perdikaris, P., and Karniadakis, G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 2019. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Ren, X., Yang, T., Wang, Y., and Zeng, W. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *ICLR*, 2022. URL https://arxiv.org/abs/2102.10543.

Richter-Powell, J., Lipman, Y., and Chen, R. T. Neural conservation laws: A divergence-free perspective. *NeurIPS*, 2022. URL https://arxiv.org/abs/2210.01741.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017. URL https://arxiv.org/abs/1609.03890.

Sato, T. K., Nauhaus, I., and Carandini, M. Traveling waves in visual cortex. *Neuron*, 2012. URL https://www.sciencedirect.com/science/article/pii/S0896627312005910.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *ICML*, 2021. URL https://arxiv.org/abs/2102.09844.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. URL https://arxiv.org/abs/2007.06600.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. URL https://arxiv.org/abs/1907.10786.

Shi, Y., Yang, X., Wan, Y., and Shen, X. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *CVPR*, 2022. URL http://128.84.21.203/abs/2112.02236.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *ICLR*, 2021a. URL https://arxiv.org/abs/2010.02502.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021b. URL https://arxiv.org/abs/2011.13456.

Song, Y., Sebe, N., and Wang, W. Orthogonal svd covariance conditioning and latent disentanglement. *IEEE T-PAMI*, 2022. URL https://arxiv.org/abs/2212.05599.

Spingarn-Eliezer, N., Banner, R., and Michaeli, T. Gan" steerability" without optimization. *ICLR*, 2021. URL https://arxiv.org/abs/2012.05328.

Tonnaer, L., Rey, L. A. P., Menkovski, V., Holenderski, M., and Portegies, J. W. Quantifying and learning linear symmetry-based disentanglement. *arXiv preprint arXiv:2011.06070*, 2020. URL https://arxiv.org/abs/2011.06070.

Toth, P., Rezende, D. J., Jaegle, A., Racanière, S., Botev, A., and Higgins, I. Hamiltonian generative networks. *ICLR*, 2020. URL https://arxiv.org/abs/1909.13789.

Tzelepis, C., Tzimiropoulos, G., and Patras, I. WarpedGANSpace: Finding non-linear rbf paths in GAN latent space. In *ICCV*, 2021. URL https://arxiv.org/abs/2109.13357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 2017. URL https://arxiv.org/abs/1706.03762.

Voynov, A. and Babenko, A. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020. URL https://arxiv.org/abs/2002.03754.

Wei, Y., Shi, Y., Liu, X., Ji, Z., Gao, Y., Wu, Z., and Zuo, W. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *ICCV*, 2021. URL https://arxiv.org/abs/2108.07668.

Zhang, R. Making convolutional networks shift-invariant again. In *ICML*, 2019. URL https://arxiv.org/abs/1904.11486.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. S3fd: Single shot scale-invariant face detector.

In *ICCV*, 2017. URL https://arxiv.org/abs/1708.05237.

Zhu, J., Feng, R., Shen, Y., Zhao, D., Zha, Z.-J., Zhou, J., and Chen, Q. Low-rank subspaces in gans. *NeurIPS*, 2021. URL https://arxiv.org/abs/2106.04488.

Zhu, J., Shen, Y., Xu, Y., Zhao, D., and Chen, Q. Region-based semantic factorization in gans. *ICML*, 2022. URL https://arxiv.org/abs/2202.09649.

Zhu, X., Xu, C., and Tao, D. Learning disentangled representations with latent variation predictability. In *ECCV*, 2020. URL https://arxiv.org/abs/2007.12885.

A. Appendix

A.1. Implementation Details

VP Score. The dataset of image pairs $[\boldsymbol{x}_0, \boldsymbol{x}_T]$ is created by randomly sampling from different interpretable directions. Since the used models have a different number of directions, the crafted datasets also have a different number of images accordingly. Specifically, the dataset consists of $10,000$ images for SNGAN and VAEs, $20,000$ images for BigGAN, and $40,000$ images for StyleGAN2. We randomly select 10% of the images as the training set and the rest as the test set. The simple neural network for the VP score evaluation consists of four stacked convolutional layers with batch normalization and ReLU activations. The learning rate is set to 0.005, and we train the network for 300 epochs with the batch size set as 32. We report the classification accuracy $(\%)$ on the test set as the score.

Pre-trained GANs. We set the total timestep T to 10 for all the datasets and models. In line with Tzelepis et al. (2021), the number of potential functions (traversal paths) K is set as 64 for SNGAN, 120 for BigGAN, and 200 for StyleGAN2. The output images are of size $64{\times}64$ for SNGAN, of size $256{\times}256$ for BigGAN, and of size $1024{\times}1024$ for StyleGAN2. During the inference stage, we also negatively traverse the latent space by $\boldsymbol{z}_t - \nabla_{\boldsymbol{z}} u^k(\boldsymbol{z}_t, t)$. The anti-symmetry of the traversal is thus achieved.

Pre-trained VAEs. We set the number of traversal path K to 32 and define the total timestep T as 10 for both MNIST and Dsprites. The training process lasts $100,000$ iterations.

Integrating Traversal into VAE Training. For MNIST, we define 3 factors of variations, *i.e.,* scaling, rotation, and color transformations. Each transformation has 8 states of variations. For Dsprites, we use the self-contained 5 factors of variations, *i.e.,* x position, y position, scaling, orientation, and shape transformations. The training also lasts $100,000$ iterations for both datasets. For the comparison fairness, the naively trained baseline employs the loss $\mathbb{E}_{\boldsymbol{z}_t}[-\log p_\theta(\boldsymbol{x}_t|\boldsymbol{z}_t) + \mathrm{D}_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}_t|\boldsymbol{x}_t)||p_{\mathcal{Z}}(\boldsymbol{z}_t)\right]$ to optimize the ELBO of the same transformed input data.

A.2. Impact of Different Losses

Table 6. Impact of different loss terms on the VP scores (%).

Models	Ours	w/o $\mathcal{L}_J$	w/o $\mathcal{L}_f$	w/o $\mathcal{L}_u$
SNGAN	**65.89**	55.37	45.78	63.19
BigGAN	**15.29**	13.91	12.87	14.68
StyleGAN2	**48.54**	41.77	36.91	46.24

Table 6 presents the complete ablation studies of losses on all the datasets. As can be seen above, when $\mathcal{L}_J$, $\mathcal{L}_f$, or $\mathcal{L}_u$ are not applied, our model would have performance degradation of different extents. The Jacobian regularization $\mathcal{L}_J$ can encourage that the trajectory could cause meaningful variations, while the PDE constraints $\mathcal{L}_f$ ensures that the potential flow follows wave-like spatial-temporal dynamics. The initial condition constraint can improve the score slightly but more importantly it is applied to help generate smoother traversal paths.

A.3. Why We Need PDE Constraints

We add the PDE constraints to the velocity fields to learn good spatial-temporal dynamics for smooth, continuous, and flexible latent trajectories. The formulation matches the space dynamics ∇u to the time dynamics $\partial_t u$, leading to stable potential flows and smooth wave-like paths in the latent space. Since the latent code is progressively updated by $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \nabla_{\boldsymbol{z}} u^k(\boldsymbol{z}_t, t)$, if no constraints are applied on the gradient, the magnitude of $\nabla_{\boldsymbol{z}} u^k(\boldsymbol{z}_t, t)$ might gradually get amplified and then eventually the latent code $\boldsymbol{z}$ is likely to go out of the manifold. Enforcing PDE constraints in spatiotemporal form could help to limit the magnitude of the gradient and create wave-like plausible trajectories.

A.4. Visual Gallery of Identified Semantic Attributes

SNGAN and StyleGAN2. Fig. 8 displays some more semantic attributes identified by our potential PDEs on SNGAN and StyleGAN2. Our method can precisely control the target image attributes while keeping other traits uninfluenced.

BigGAN. Previous disentanglement approaches heavily rely on human faces and animal images for visualization. Here we

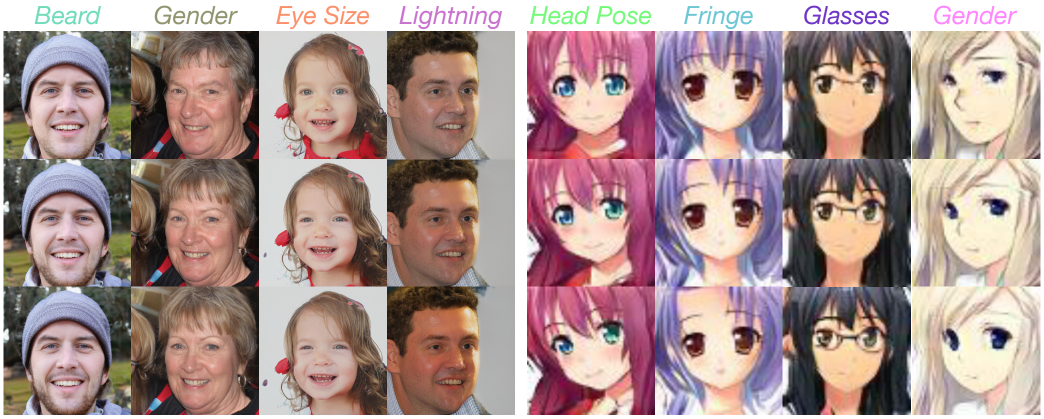

Figure 8. More semantics discovered by our learned potential PDEs on SNGAN and StyleGAN2.

instead show some results with alternative objects belonging to the ImageNet classes based on BigGAN. Fig. 9 presents such traversal results. Our potential PDEs are still able to identify distinct semantics from images of various categories.

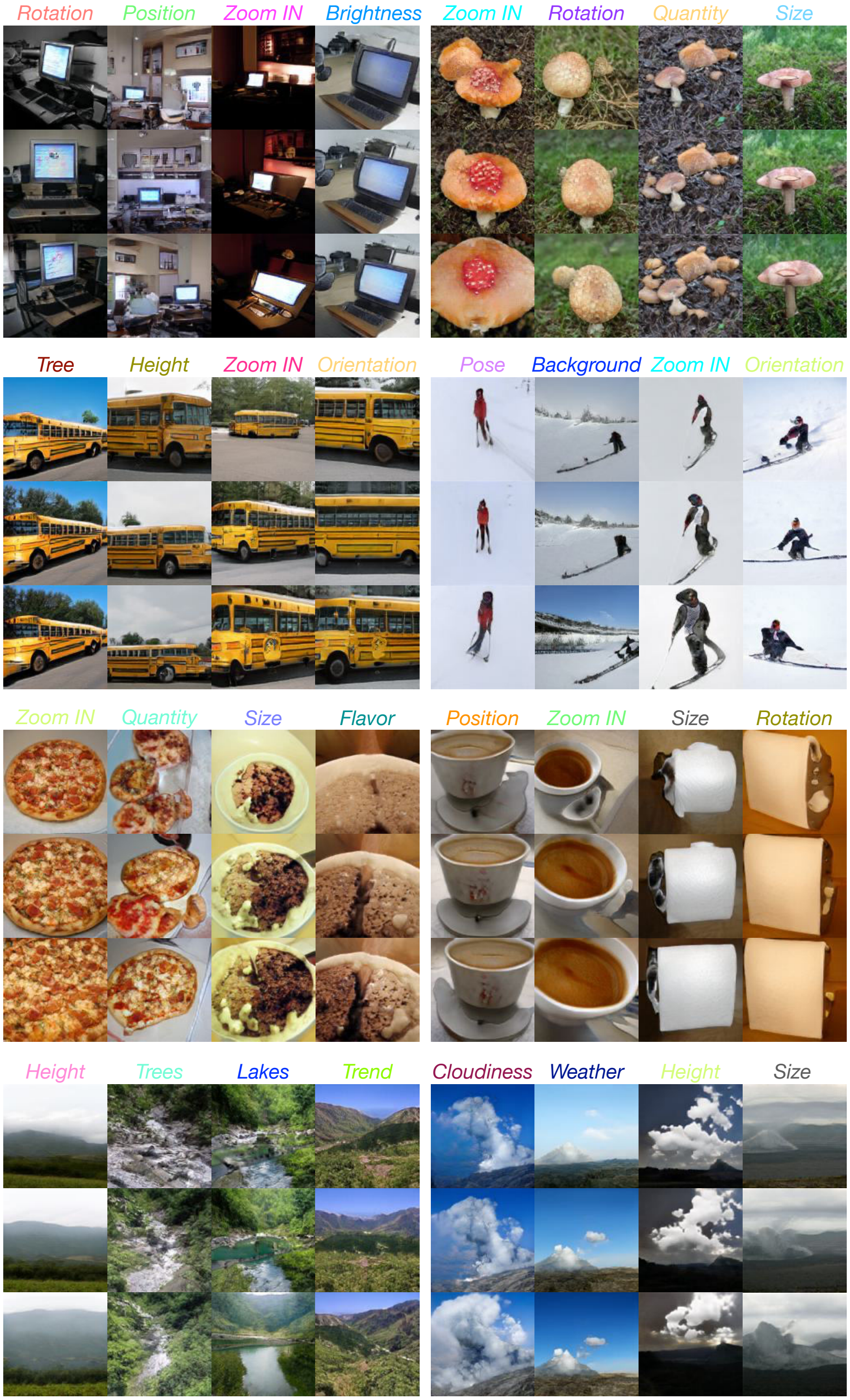

Figure 9. Semantic attributes of different objects discovered by our learned potential PDEs on BigGAN. The specific image categories include Computer Screen (1_{st} row left), Mushroom (1_{st} row right), Schoolbus (2_{nd} row left), Ski (2_{nd} row right), Pizza and Ice Cream (3_{rd} row left), Coffee and Toilet Tissue (3_{rd} row right), Valley (4_{th} row left), and Volcano (4_{th} row right).