MSc Artificial Intelligence
Master Thesis

Physics-Informed Representation Alignment

by
Antonios Tragoudaras
15233421

36 EC
Jan. 2025 - Sep.2025

Daily Supervisor:
Dr A Zadaianchuk

Supervisor/Examiner:
Associate Prof. E. Gavves

ELLIS MSc Honors Supervisor:
Assistant Prof. F. Locatello

Second reader:
MSc D. Cherniavskii

Universiteit van Amsterdam

Contents

ABSTRACT

Envisioning large-scale Video Generative Models (VGMs) as world simulators represents a significant frontier in Artificial Intelligence, promising to empower the next generation of Physical AI; enabling embodied agents to learn, plan, and simulate actions in a safe, scalable digital twin of our physical world. Nevertheless, the realization of this vision is hindered by the models' limited understanding of physics. Concurrent works have revealed that these models have only developed immature physics reasoning capabilities, as an emerging from their generative pre-training on massive, unstructured video datasets. The aggregated knowledge is a fragile imitation of visual pattern visual patterns present in the training data, rather than a truly grasp of the underlying physical dynamics. Thus, despite their unprecedented visual fidelity abilities in generating videos, these models frequently defy fundamental physical laws. Existing methods struggle to bridge this gap: imposing explicit control at inference time does not enhance the model's intrinsic knowledge, while prior knowledge distillation methods via representation alignment relies on opaque, black-box vision encoders, suffers from training instabilities.

To address these limitations, we introduce **P**hysics-**I**nformed **R**epresentation **A**lignment (PIRA), a framework for instilling targeted, interpretable physical knowledge into pre-trained Video Diffusion Models. Our approach is based on distilling knowledge from physics-rich proxy signals—representations of the observable consequences of physical laws, such as an optical flow field, relative depths, segmentation masks serving as a proxy of an object's state variable. This is a scalable approach for teaching simple motions that adhere to Newtonian Dynamic laws. In our work we focus on items falling under normal gravity. The core of our design is to re-purpose the VDM's native VAE encoder to create inherently compatible *teacher* representations from these signals. Developing PIRA also necessitated a more principled evaluation of physical plausibility. We identify that existing benchmarks suffer from fundamental flaws, such as the subjectivity of Vision Question Answering based scores or the false negatives produced by single-outcome trajectory matching. We therefore introduce a novel, evaluation strategy that moves beyond these limitations by measuring a generated video's adherence to governing dynamical equations and conservation of physical invariants. Through extensive experiments, our findings reveal that PIRA is highly effective at teaching Video Diffusion Models to respect underlying physical principles. This work marks a significant step toward grounding Video Diffusion Models in some form of causal principles of the physical world, enhancing their reliability and trustworthiness as world simulators. All data, weights, and code are open-sourced at: https://github.com/physics-informed-REPA.

Chapter 1

Introduction

A central ambition in modern artificial intelligence is the creation of Cyber-Physical AI systems characterized with a human-level intuitive grasp of the physical world that enables robust embodied reasoning (Liu et al., 2025; Xiang et al., 2025). In pursuit of this goal, two prominent, yet philosophically apart, paradigms have emerged. The *embodied-action* paradigm centers on Vision-Language-Action (VLA) models. These systems, exemplified by recent generalist models, like π_0, GROOT N1, and Gemini Robotics (Black et al., 2024; Bjorck et al., 2025; Team et al., 2025), have demonstrated dexterity in robotics task. Such models are explicitly trained to map their multi-modal sensory inputs to continuous robot actions. Learning from vast, specialized datasets of expert demonstrations, their primary objective is to achieve deft control in robotic manipulation-like tasks. In parallel, an *observational-learning* paradigm has gained tremendous momentum through large-scale Video Generative Models (VGMs). Unlike VLAs trained for direct control, these models are built on a simple yet immensely scalable generative objective—such as reversing a gradual noising process in diffusion models Ho et al. (2020), applied to massive corpora of passive, "in-the-wild" video. Their training objective is to synthesize coherent future frames given some context(usually visual/textual or both).

The astonishing success of this *learning by watching* approach has led their creators to envision a far more promising usage case. Early works (Brooks et al., 2024; Agarwal et al., 2025) envisioned that the constant scaling of VGMs is a promising path towards the development of capable simulators of the physical world we live in. This ambitious vision is substantiated by a fascinating consequence of their training: state-of-the-art VGMs are becoming potent **zero-shot reasoners**, capable of solving a diverse array of visual tasks for which they were never explicitly trained (Wiedemer et al., 2025). This emergent phenomenon suggests that in order to successfully generate realistic video evolutions (unrolling of future), the models are forced to implicitly learn the underlying dynamics of the world. Most notably, this includes an emerging capacity for modeling intuitive physics.

In this work, we systematically explore the observational-learning paradigm to critically assess the depth and reliability of the physical knowledge that emerges from it. While the emergent capacity for intuitive physics is promising, it raises a fundamental question: is this implicitly learned knowledge grounded in the causal, dynamical principles of the physical world, or is it a fragile, superficial mimicry of visual patterns observed during training? This question becomes critical because, despite their unprecedented visual fidelity, a closer examination reveals that their capability for generating physically plausible content remains limited. An emerging line of research has identified a critical *physics gap*, revealing that even state-of-the-art models frequently produce videos with dynamics that violate fundamental physical laws (Zhang et al., 2025a; Motamed et al., 2025; Bansal et al., 2024; 2025; Gu et al., 2025; Meng et al., 2024). Existing methods fail to bridge this gap as they either impose explicit control at inference time (Zhang et al., 2025c; Burgert et al., 2025), or leverage knowledge distillation from black-box semantic models (Zhang et al., 2025b). The former does not enhance the model's intrinsic understanding, while the latter can be unstable and lack interpretability, liable to training instabilities and catastrophic forgetting during supervised fine-tuning (SFT).

To address these limitations, we introduce **P**hysics-**I**nformed **R**epresentation **A**lignment (PIRA), a framework for distilling physical knowledge into pre-trained Video Diffusion Models.Our approach is founded on the principle that while directly encoding physical laws is complex, we can provide a powerful learning signal by using representations of their *observable consequences*. For instance, the motion of an object under gravity is governed by Newtonian dynamics; a dense optical flow field capturing that motion serves as an explicit, frame-by-frame proxy for the object's velocity field, a core state variable in those dynamics. The core of our

contribution is a simple yet elegant architectural design: we re-purpose the VDM's native *native VAE* of our Video Diffusion Model —the component responsible for tokenizing video into a latent representation— as a versatile feature extractor for our *teacher* signals. By feeding this encoder physics-rich proxy signals, such as object segmentation masks, depth maps, or optical flow fields, we construct teacher representations that are inherently compatible with the VDM's noisy hidden states. This design bypasses the latent space mismatch issues described in prior work (Zhang et al., 2025b), which often lead to training instability. We pair this architecture with a robust representation alignment objective. Instead of enforcing a direct, feature-by-feature matching process that can be brittle when aligning representations from different distributions (clean physics proxies vs. noisy video latents), our objective aligns the overall relational structure of the feature spaces, encouraging their pairwise similarity matrices to correspond. Using this stable alignment strategy, we instill targeted physical knowledge into the VDM on a carefully crafted synthetic dataset of 10k videos designed to capture fundamental free-fall dynamics.

Developing such a framework, however, required us to co-design a more principled way to measure physical plausibility in the first place. Current evaluation strategies largely follow two main paradigms. One line of work leverages Vision-Language Models (VLMs) or human annotators to provide qualitative judgments on "physical commonsense," as seen in benchmarks like VideoPhy (Bansal et al., 2024; 2025). A second approach involves quantitative trajectory matching, where a generated object's motion is compared against a single ground-truth trajectory from a simulation or recording (Li et al., 2025; Motamed et al., 2025). A comprehensive review of these strategies is provided in Chapter 2, but both suffer from fundamental limitations: the former is prone to subjectivity and VLM hallucination, while the latter can produce false negatives by unfairly penalizing a physically plausible video that simply follows a different valid trajectory. To overcome these shortcomings, we introduce a novel, two-fold evaluation strategy that moves beyond subjective scoring and single-outcome comparisons. We not only implicitly measure fidelity to a reference but also explicitly quantify a video's adherence to the governing laws of motion and its conservation of physical invariants, providing a more robust measure of a model's true physical understanding.

To this end this work makes three primary contributions toward creating more physically-grounded video generative models:

1. We propose PIRA, the first representation alignment framework explicitly designed for distilling targeted, interpretable physical knowledge into pre-trained VDMs. We demonstrate its flexibility by successfully using diverse physics-rich proxies (masks, depth, and optical flow) as teacher signals (Section 3).
2. We introduce a comprehensive, two-fold evaluation strategy that moves beyond simple trajectory matching. On that end, we quantify physical plausibility both implicitly, by matching generated state variables (e.g., trajectory, object permanence) against a ground-truth reference, and explicitly, through a novel suite of physics-informed scores that measure adherence to governing dynamical equations and conservation of physical invariants.
3. Through extensive experimentation, we demonstrate that PIRA significantly outperforms standard fine-tuning and naive alignment-based methods across all metrics. Our results reveal that PIRA is effective at teaching Video Diffusion Models (VDMs) to respect physical dynamics, marking a substantial step towards developing more reliable and trustworthy world simulators.

The rest of this work is organized as follows. Chapter 2 reviews the background, covering the role of Video Generative Models as world simulators, existing approaches to physics-aware generation, and methods for quantifying physical reasoning in Video Diffusion Models. Chapter 2.2 provides the necessary technical preliminaries on diffusion models and representation alignment. Chapter 3 introduces our proposed framework, PIRA, detailing its high-level architectural design and the specific alignment objectives used to distill physical knowledge. Chapter 4 presents Morpheus, our novel benchmark for evaluating explicit physical reasoning, which is designed to address a critical gap in the field of Physical AI, which we try to fill. Chapter 5 motivates our methodology for modeling free-fall dynamics and our strategy for building a high-quality synthetic dataset that enables principled evaluation of physics. In Chapter 6, we present our comprehensive experimental setup, implementation details, and a detailed analysis of our quantitative and qualitative results. Finally, Chapter 7 summarizes our contributions and discusses their implications, while Chapter 8 outlines the limitations of our methodology and proposes directions for future research.

Chapter 2

Background

This chapter establishes the foundational context for our work. We begin by reviewing the relevant background and related research in physics-aware video generation, covering the evolution of Video Generative Models (VGMs) as world simulators, methods for evaluating physical plausibility, and existing strategies for instilling physics into generative models. Subsequently, we provide the necessary technical preliminaries, detailing the core mechanics of Latent Diffusion Models and the principles of representation alignment that underpin our proposed framework. This consolidated chapter is designed to equip the novice reader with both the conceptual and technical tools required to understand our contributions, while allowing those familiar with the topics to proceed directly to our proposed methodology (Chapter 3).

2.1 Related Work

The pursuit of physically plausible video generation exists at the intersection of several active research domains. To position our work clearly, we first review the broader vision of Video Generative Models (VGMs) as world simulators and the critical challenge posed by their frequent physical inconsistencies, ofter framed as the gap in *physical reasoning*. We then review the concurrent efforts to quantify this gap through various evaluation benchmarks. Finally, we examine and distinguish between the two dominant philosophies for instilling physical plausibility into VGMs: explicit, inference-time control and knowledge distillation during training. This final area, particularly distillation via representation alignment, serves as the direct precursor to our own contributions.

2.1.1 Video Generative Models as World Simulators

The concept of learning a predictive model of the world has long been a central goal in artificial intelligence. A seminal work by Ha & Schmidhuber (2018) introduced the modern paradigm of a *world model*: a generative neural network that learns a compressed spatio-temporal representation of an environment, serving as a rapid internal simulator for an agent to learn policies. In recent years, this vision has scaled dramatically. Flagship Video Generative Models (VGMs) like OpenAI's Sora (Brooks et al., 2024), Google's Veo (Google, 2024), and NVIDIA's Cosmos (Agarwal et al., 2025) have demonstrated a remarkable ability to generate coherent, high-fidelity dynamic scenes. This success has led to their reframing as general-purpose **World Foundation Models (WFMs)** a digital twin of the physical world intended to power the next generation of Physical AI.

The central promise of a WFM is to serve as a safe and scalable environment where an embodied agent, such as a robot, can learn, plan, and simulate the consequences of its actions. However, despite their impressive visual fidelity, a critical limitation prevents these models from serving as reliable world simulators: their frequent failure to generate physics-plausible future generations that adherence to fundamental physical laws. The discrepancy between visual fidelity and intuitive physical understanding constitutes a significant gap in *physics-modeling*. This shortcoming is not merely anecdotal; a growing body of research (Zhang et al., 2025a; Bansal et al., 2024; 2025; Motamed et al., 2025; Meng et al., 2024; Chow et al., 2025) has rigorously documented it. This discrepancy, often termed the "physics gap," has given rise to two complementary philosophical approaches to bridge it.

One of the current popular approach is to treat physical understanding as an *emergent property of scale*. This data-driven philosophy posits that by training massive Vision-Language-Action (VLA) models on vast

and diverse datasets of embodied interaction, a functional understanding of physics will implicitly emerge. State-of-the-art models like π_0 (Black et al., 2024), $\pi_{0.5}$ (Intelligence et al., 2025), Gemini Robotics (Team et al., 2025; Abdolmaleki et al., 2025), and GROOT N1 (Bjorck et al., 2025) exemplify this paradigm. These models leverage powerful Vision-Language Model (VLM) backbones to interpret multi-modal inputs and are trained on heterogeneous "data pyramids" that combine real-robot data, human videos, and simulation. Architecturally, they achieve precise and fluent manipulation by delegating continuous action generation to a separate module, often termed an *action expert* (Black et al., 2024). This expert is typically a smaller transformer trained with a diffusion-based objective, such as the flow matching (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023) used in π_0 and GROOT N1, to produce smooth, high-frequency action chunks. This design allows VLAs to tackle a wide range of dexterous manipulation tasks, from long-horizon taks like folding laundry and cleaning tables (Team et al., 2025; Intelligence et al., 2025) to precise skills evaluated on benchmarks like the Open X-Embodiment dataset (O'Neill et al., 2024), which is a cornerstone for training these generalist policies.

The second approach, which our work belongs to, argues for the *explicit integration of physical principles*. Instead of hoping for physics to emerge, this paradigm seeks to instill it directly into the generative process. This can be achieved through explicit physical simulation during generation (PAG-E), or via implicit learning (PAG-I) where the consequences of physical laws guide training (Liu et al., 2025). Our work, PIRA, is a prime example of PAG-I. By using first-principles physics proxies to regularize the internal representations of a pre-trained VDM, we are not just fine-tuning for a task; we are explicitly teaching the model to respect the underlying causal structure of the physical world. This aligns with the vision of transforming powerful but physically-naive VDMs into true, reliable World Foundation Models (Agarwal et al., 2025) by grounding their generation process in the laws of physics, thereby closing the critical *physics gap* that currently limits their applicability as trustworthy world simulators.

2.1.2 Evaluation of Physical Plausibility and Reasoning

A key challenge in developing physics-aware generative models is the capability of benchmarking their physical reasoning in the first place. A significant body of work has emerged to quantify the *physics gap*, largely following two main paradigms, summarized in Figure 2.1.

The first approach leverages Vision-Language Models (VLMs) to provide qualitative judgments on "physical commonsense". Benchmarks like VideoPhy and VideoPhy2, (Bansal et al., 2024; 2025), and PhyGenBench (Meng et al., 2024) prompt VDMs to generate videos depicting specific physical phenomena, and human evaluators and VLM experts to derive a Vision Question Answering (VQA) score for the outcome. As illustrated in Figure 2.1(a), while helpful in assessing high-level semantic adherence, this paradigm is prone to the inherent limitations of VLMs, including subjectivity, hallucination (Li et al., 2023), while usually fails to capture nuanced physical inconsistencies (Chow et al., 2025), failing to provide the objective, quantitative evidence required for rigorous scientific assessment.

The second approach involves quantitative trajectory matching, where the generated motion of an object is compared against a single ground-truth trajectory from a simulation or a real-world recording (Li et al., 2025; Motamed et al., 2025; Agarwal et al., 2025). This method, however, suffers from a critical flaw: it can produce false negatives by unfairly penalizing a generated video that can be physically plausible but simply follows a different trajectory due to unobserved initial conditions (e.g., slight variations in friction, mass, or initial velocity). This limitation makes trajectory matching an unreliable measure of a model's innate physical understanding.

The shortcomings of both VLM-based and trajectory-based evaluations highlight the need for a more principled framework that moves beyond subjective scoring and single-outcome comparisons. In this study, we address this need by co-designing a novel evaluation methodology in Section 4 that directly measures generated video's adherence to physical invariant and underlying governing equations (Figure 2.1). Chapter 4 covers our evaluation startegy comperhensively.

2.1.3 Strategies for Improving Physical Plausibility: Control vs. Distillation

In response to the identified physics gap, the research community has explored several strategies to enhance the physical realism of video generation. These efforts can be broadly categorized into two distinct philosophies: those that explicitly intervene during the generation process at sampling time, providing extra levels of control, and those that try to distill physical understanding into the model's parameters during training, with a Supervised-Finetuning-based (SFT) process.

Explicit Control as a Pathway to Plausibility

A major line of research treats VGMs as high-fidelity renderers that can be commanded by explicit, fine-grained control signals(Zhang et al., 2025c; Burgert et al., 2025). While often demonstrated with user-defined inputs, these methods fundamentally offer a powerful form of **composability**: they decouple the synthesis of photorealistic video from the generation of motion. This allows the VDM to be guided by any external model that can produce a motion plan, effectively outsourcing the task of physical plausibility.

Prominent approaches in this category operate on different control modalities. **Trajectory-based control**, exemplified by *Tora* (Zhang et al., 2025c), introduces modules that fuse precise motion paths directly into the diffusion model's generation process, forcing objects or the camera to follow a given trajectory. A different philosophy, seen in *Go-with-the-Flow* (Burgert et al., 2025), achieves similar control by structuring the initial noise of the diffusion process according to an optical flow field. In both cases, if the input trajectory or flow field is generated not by a human but by an external physics engine or a dynamics prediction model, the VDM is de-facto constrained to generate a plausible outcome.

Another form of explicit guidance comes from dense **spatial conditioning**. The influential work on *ControlNet* (Zhang et al., 2023) demonstrated how conditioning signals like depth maps, normal maps, or object skeletons can be used to guide image generation. This principle naturally extends to video, where a sequence of physically consistent depth maps or 3D poses from a simulator can enforce properties like object rigidity, perspective, and coherent motion over time(Alhaija et al., 2025).

Despite highly effective, these control-centric approaches fundamentally *command* the VGM, rather than *teach* it physics. A a result the model's own intrinsic understanding of physics does not improve. The burden of plausibility is outsourced to the external guidance model. This reliance on an external component for physical guidance motivates an alternative strategy, which our work pursues: distilling physical knowledge directly into the generative model's parameters.

Knowledge Distillation via Representation Alignment

A second, distinct philosophy aims to bake a more fundamental understanding of the world directly into the model's parameter space. This is achieved by regularizing the model's internal representations to align with those of a well-informed *teacher* model during training. This paradigm shifts the objective from controlling a single output to teaching the model how to behave more like an expert.

The seminal work in *REP*resentation *A*lignment (REPA)(Yu et al., 2024), showcased that aligning the intermediate features of an image diffusion model Ma et al. (2024); Peebles & Xie (2023) with those from a powerful Vision Foundation Model (VFM) like (Oquab et al., 2023) could dramatically accelerate training from scratch. This was achieved via a direct cosine similarity objective that explicitly maximized the feature similarity between student and teacher features.

VideoREPA(Zhang et al., 2025b) extends the representation alignment principle to improve pre-trained video diffusion models. While VideoREPA successfully demonstrated that relational alignment can transfer knowledge from a VFM into a VDM, its teacher remains a black-box model, providing a powerful but ultimately correlational signal, especially for physics understanding. This leaves an open question, which

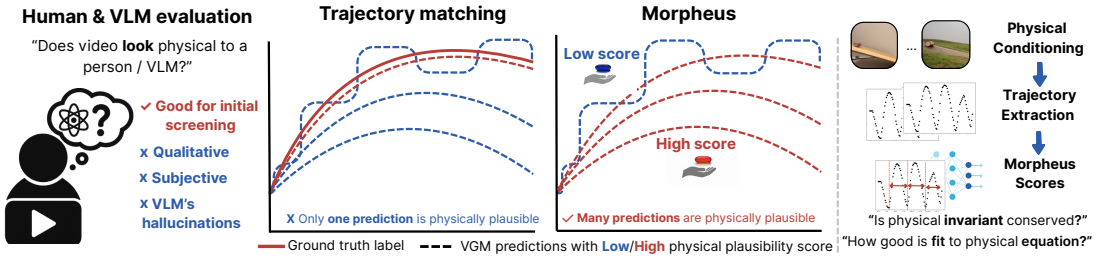


Figure 2.1: Comparison of evaluation methods for physical plausibility in video generation, adapted from (Zhang et al., 2025a). **(a)** Human or VLM-based judgments provide qualitative and subjective assessments. **(b)** Trajectory matching compares a generated path to a single ground-truth, potentially miss-classifying other physically valid trajectories as incorrect. **(c)** Our unbiased evaluation approach, based on (Zhang et al., 2025a), moves beyond single outcomes to assess adherence to fundamental physical laws by evaluating the conservation of invariants and consistency with governing equations via physics-infrormed scores.

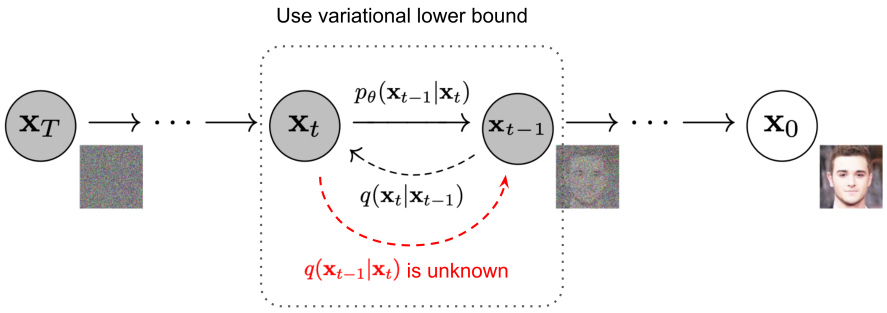


Figure 2.2: The forward (noising) and reverse (denoising) Markov chains in a Diffusion Probabilistic Model. The model learns to reverse the diffusion process to generate a sample from pure noise. Image adapted from (Ho et al., 2020).

we address in our work: can we design a more interpretable, minimal *teacher*to distill a more *causal* understanding of physical laws?

2.2 Technical Preliminaries

In this section, we cover the relevant technical preliminaries required to understand our proposed framework. We begin by demystifying the inner workings of diffusion in probabilistic molding, the foundational concepts of Latent Diffusion Models, and the architectural shift from UNet backbones to modern Transformer-based architectures. Subsequently, we present the representation alignment methods, which lay down the theoretical basis for envisioning and manifesting our physics-informed alignment paradigm.

2.2.1 Latent Diffusion Models & Diffusion Transformers

Latent Diffusion Models Diffusion Probabilistic Models (Ho et al., 2020) belong to one of the most powerful classes of generative models, that learn to synthesize data by reversing a gradual noising process, achieving unprecedented quality in visual image synthesis, surpassing the previous gold-standard of GANs (Goodfellow et al., 2014). The diffusion process can be defined with a forward Markov-Chain that incrementally adds Gaussian noise to an input sample $\mathbf{x}_0$ over T timesteps, producing a sequence of noisy latents $\mathbf{x}_1, \ldots, \mathbf{x}_T$ (forwrd process). The key idea is to learn the inverse process $\mathbf{p}_\theta$ parameterized with a neural network, which is trained with the objective to denoise $\mathbf{x}_t$ to predict a less noisy $\mathbf{x}_{t-1}$, as illustrated in Figure 2.2.

While effective, applying this process directly in pixel space is not efficient for high-resolution data, like images and videos. *Latent Diffusion Models (LDMs)* (Rombach et al., 2022) tackle the latter challenge by operating in the compressed latent space of a pre-trained Variational Autoencoder (VAE) (Kingma & Welling, 2013; 2019). The encoder $\mathcal{E}$ network of VAE first maps a high-dimensional image $\mathbf{x}_0$ to a lower-dimensional latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. The forward diffusion process is then normally carried out with latent variable $\mathbf{z}_0$. The denoising network is trained to predict the noise ϵ added to the latent at timestep t. This prediction is denoted with ϵ_{theta}. Using a standard ϵ-parametrization, the simplified objective can be written as:

$$L_{\text{LDM}} = \mathbb{E}_{t\sim\mathcal{U}(1,T),\mathbf{z}_0,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right] \tag{2.1}$$

where $\bar{\alpha}_t$ are pre-defined coefficients of the noise scheduler. During sampling, a sample is generated by iteratively denoising a random latent variable $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and then mapping the final denoised latent $\mathbf{z}_0$ back to pixel space using the VAE's decoder $\mathcal{D}(\mathbf{z}_0)$.

From UNet to Diffusion Transformers The conventional choice for the denoising network $p\theta$ in LDMs has been the *UNet architecture* (Ronneberger et al., 2015). Its hierarchical, convolutional structure with skip connections is highly effective at capturing multi-scale spatial features. However, research has shown that the performance of U-Net blocks does not scale as predictably with increasing model size or computational resources compared to Transformers (Vaswani et al., 2017).

A pivotal shift in the field was the introduction of the *Diffusion Transformer (DiT)* (Peebles & Xie, 2023), which replaces the convolutional UNet backbone by finding a proper way to integrate the standard Transformer architecture in the LDM logic. In the DiT framework, the noisy latent $\mathbf{z}_t$ is first broken into a

sequence of patches, which are then treated as tokens, similar to the Vision Transformer (ViT) patchification process Dosovitskiy et al. (2020). The Transformer operates on this sequence of spatio-temporal tokens, conditioned on the timestep t and any other context (e.g., text embeddings, class labels, etc) via adaptive normalization layers or cross-attention. This architectural change proved to be a breakthrough for two key reasons:

- **Scalability:** DiTs exhibit remarkable scaling properties. As demonstrated in their founding paper, increasing model size (depth/width) or training compute directly translates to improved generative performance (e.g., lower FID scores), a hallmark of the Transformer architecture.
- **Architectural Simplicity and Versatility:** By leveraging a standard Transformer, DiTs benefit from the vast ecosystem of optimizations and innovations from natural language processing, simplifying the handling of various conditioning modalities.

The success of this scalable paradigm paved the way for its application in large-scale video generation models like Sora (Brooks et al., 2024) and CogVideoX (Yang et al., 2024), the latter of which serves as the backbone model for our work.

Video Diffusion Model (VDM) Backbone for Video Generation Our work utilizes the CogVideoX (Yang et al., 2024) model as its backbone, an open-source DiT designed for high-quality, long-duration video generation. The CogVideoX architecture introduces several key innovations. It replaces the standard 2D image VAE with a powerful, natively trained *3D causal VAE*. The VAE (Kingma & Welling, 2013; 2019) compresses videos across both spatial and temporal dimensions simultaneously, significantly reducing sequence length and helping to prevent the flickering artifacts common in methods that finetune 2D VAEs. Second, it substitutes the factorized space-time attention of earlier VDM with a *3D full attention* mechanism within its Transformer blocks, allowing every patch to attend to every other patch across both space and time. This holistic approach is better suited for capturing complex, large-scale motion. Finally, it employs an *Expert Transformer* design with expert adaptive LayerNorm, which facilitates a deeper and more effective fusion of textual and visual modalities. This combination of a powerful 3D tokenizer and a scalable, fully spatio-temporal Transformer provides a powerful VDM.

2.2.2 Representation Alignment

Representation alignment has recently emerged as a powerful regularization technique for training diffusion models more efficiently. The core idea is to guide the internal representations of a diffusion model to align with features extracted from a powerful, pre-trained Vision Foundation Model (VFM), such as DINOv2

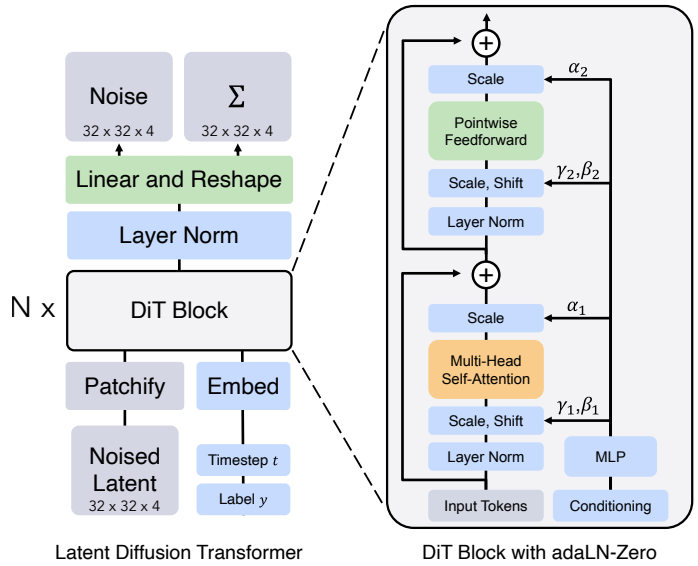


Figure 2.3: The architectural block of the **Diffusion Transformer (DiT)**. Timestep and other conditioning information are injected via an adaptive layer normalization (adaLN-Zero) block. Figure is taken from (Peebles & Xie, 2023).

(Oquab et al., 2023). This paradigm effectively distills rich semantic knowledge from the VFM into the generative model.

The foundational work, **REPresentation Alignment (REPA)** (Yu et al., 2024), demonstrated that this approach could dramatically accelerate the training convergence of DiTs. In the REPA framework, the VFM acts as a "teacher" that provides target representations $\mathbf{y}^*$ from a clean image $\mathbf{x}$. The denoising DiT acts as a "student," whose intermediate hidden states $\mathbf{h}_t$ are derived from a noisy version of the image, $\mathbf{z}_t$. The alignment objective regularizes the training by encouraging the student's representations to match the teacher's, thereby bridging the semantic gap. Specifically, the transformer encoder hidden states $\mathbf{h}_t$ are projected with a small feed-forward layer resulting in h_ϕ , to match the dimensionality with the teacher features $\mathbf{y}^*$ by maximizing their similarity, as shown in Figure 2.4. The alignment loss is typically formulated as:

$$\mathcal{L}_{\text{REPA}} = -\mathbb{E}_{\mathbf{x},\epsilon,t}\left[\frac{1}{N}\sum_{n=1}^{N}\text{sim}(\mathbf{y}_n^*, h_\phi(\mathbf{h}_t)_n)\right], \tag{2.2}$$

where N is the number of patches and $\text{sim}(\cdot,\cdot)$ is a similarity metric, specifically the direct cosine similarity. The total loss becomes a weighted sum of the standard diffusion loss and the REPA loss: $\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda\mathcal{L}_{\text{REPA}}$.

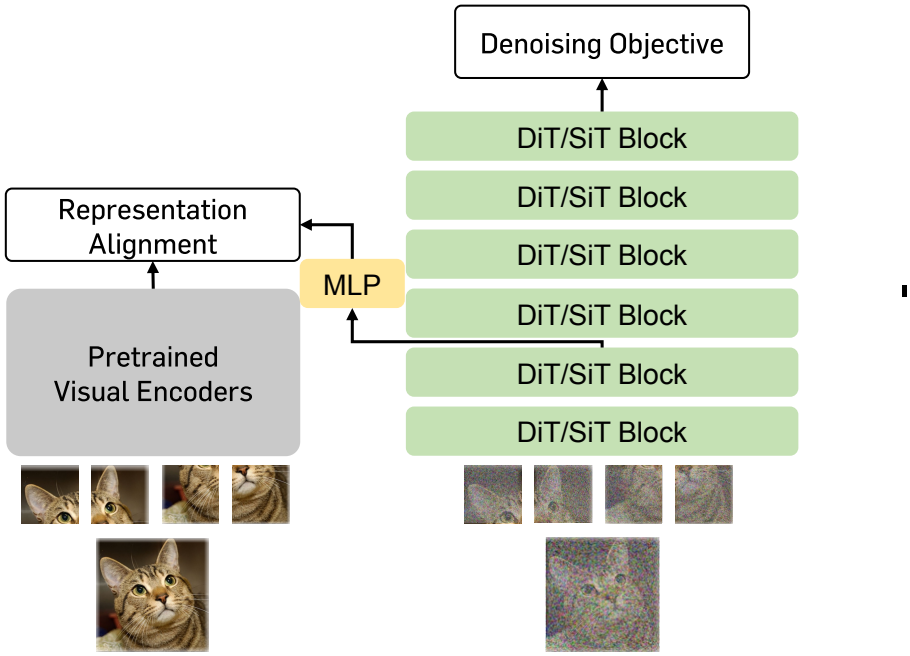


Figure 2.4: The principle of Representation Alignment (REPA). The internal representations $\mathbf{h}_t$ of a denoising transformer, which processes a noisy latent $\mathbf{z}_t$, are aligned with the rich semantic features $\mathbf{y}^*$ extracted from the corresponding clean image $\mathbf{x}$ by a pre-trained visual encoder (e.g., DINOv2). Figure is taken from (Yu et al., 2024).

Building on this, subsequent works have extended the alignment principle. *REPA-E* (Leng et al., 2025) demonstrated that the REPA loss provides a stable enough training signal to unlock end-to-end finetuning of both the VAE tokenizer and the diffusion model simultaneously, a process that fails with the standard diffusion loss alone. This joint optimization leads to even greater training acceleration.

A different but related line of work, exemplified by *VA-VAE* (Yao et al., 2025), tackles the "reconstruction vs. generation" dilemma in the space of VAE/tokenizers. The authors identify that VAEs optimized purely for pixel-perfect reconstruction can create high-dimensional latent spaces that are difficult for the diffusion model to learn, thus harming generation quality. Their proposed solution is to align the VAE's latent space directly with VFM features during the VAE's pre-training phase. By doing so, they define a more structured, *generation-friendly* latent space, which in turn makes the subsequent diffusion model training more efficient, effective, and stable.

In our work, we adapt the representation alignment principles, not for accelerating training from scratch, but for the novel task of instilling targeted, first-principles physical knowledge into a pre-trained VDM.

Chapter 3

Physics-Informed Representation Alignment

In this section, we introduce Physics-Informed Representation Alignment (PIRA), a scalable and model-agnostic paradigm for enhancing the physics understanding of pre-trained Video Diffusion Models (VDMs). Our methodology is built on the principle of representation-alignment (Yu et al., 2024), where we regularize the VDM's internal representations to align with a "teacher" that provides a form of explicit understanding of physics. Yet, our objective shifts away from accelerating the convergence of training transformer-based diffusion models (Peebles & Xie, 2023; Ma et al., 2024) from scratch and focuses instead on instilling physics knowledge into generic, already pre-trained VDMs (Yang et al., 2024). We begin by outlining the core challenges in designing such a framework, then we introduce our solution.

3.1 The Challenge of Crafting a Physics-Aware Teacher

Inducing physical knowledge into a generative model is a non-trivial task. We hypothesize that an effective way to regularize a VDM is to guide its internal representations by designing a teacher encoder that has an explicit understanding of the physical world. However, creating such a teacher presents two fundamental challenges that must be solved simultaneously.

First is the challenge of *encoding physics*. Directly encoding first-principle physical laws is an open research problem. While promising directions exist, such as using Physics-Informed Neural Networks (PINNs) (Raissi et al., 2017a;b) or Lagrangian Neural Networks (LNNs) (Cranmer et al., 2020), these approaches are often highly specialized and sensitive. Meaning they would likely require distinct model and separate pre-training phases per different phenomena thus limiting their generality. Instead, we argue that a more practical path is to create a *proxy-physics teacher* by encoding observable signals, such as motion, depth, or object shape, that are direct consequences of the underlying physical laws. We argue that latter signals can provide a coarse-mapping from object-level properties evolving over time to the underlying state variables that capture the physical dynamics and follow a specific law instantiation (like gravity).

Equally important is the creation of rich and meaningful representation alignment targets, compatible with VDM noisy hidden states. As demonstrated by recent works like VideoREPA (Zhang et al., 2025b) and CREPA (Hwang et al., 2025), extending representation alignment to pre-trained VDMs is fragile. The "hard alignment" objective of the original REPA study (Yu et al., 2024), enforces direct per-feature alignment between student and teacher features, often fails during SFT. This is because the noisy, intermediate latent states of a VDM ($\mathbf{h_t}$) live in a feature space that is fundamentally discordant compared to the clean, structured space of an external teacher encoder ($\mathbf{y}^\star$). This mismatch can cause the training dynamics to collapse, catastrophically derogating the VDM's pre-trained world knowledge. The solution, proposed by VideoREPA, is a "soft alignment" approach, which orients the relational structure of the feature spaces rather than the features themselves. The proposed solution, named Token Relational Distillation, has separate terms for promoting the temporal and spatial consistency that a VDM needs.

Therefore, to build a minimal working PIRA framework, we must solve both problems: construct a powerful encoder that can transform an explicit physics proxy into a dense, meaningful, and dense target representation,

while ensuring that this representation is compatible with the student VDM's latent space to allow for stable alignment without sacrificing any already acquired *world-knowledge*.

3.2 PIRA: A Native Encoder for Physics Distillation

A central design choice in any physics-informed alignment framework is the source of the teacher encoder. One intuitive strategy is to employ an external, *motion-specialized encoder*, similar to the Trajectory Extractor (TE) from Tora (Zhang et al., 2025c), to process physics proxies like optical flow. We treat this approach an early variant (see Appendix A.1 for details) for improving physical reasoning of VDMs via representation alignment, which proves to be functional but ultimately sub-optimal for the following critical reasons:

1. **Compromised and Uninterpretable Teacher Signal:** The Tora TE was not trained to produce a pure representation of motion. It was jointly trained with a Motion-Guidance Fuser (MGF) for task of *fusing* the the transformed motion signal with the intermediate states of VDMs. This means its representations are a compromise, not a direct encoding of the physical state, making the teacher signal ill-posed.
2. **Lack of Generality and Model Dependency:** Such specialized components are often finetuned for a specific VDM backbone, making the method difficult to generalize and tying it to a single model architecture instead of creating a universal framework.

Considering these limitations, we propose our framework: **Physics-Informed Representation Alignment (PIRA)**. As illustrated in Figure 3.1, we leverage the VDM's own native 3D-VAE as a versatile physics encoder. We hypothesize that a sufficiently powerful pre-trained VAE, designed to compress high-frequency pixel information, is more than capable of encoding lower-frequency, structured signals, such as segmentation masks, depth maps, and optical flow fields.

This elegant design choice allows us to create alignment targets $\mathbf{y}^{\star}_{\mathbf{physics}}$ that are, by definition, compatible with the VDM's latent space. This completely alleviates the need for a specialized, motion-specific encoder that requires separate pre-training and avoids the interpretability issues of the TORA-based approach. We empirically validate this claim in Section 6, where PIRA demonstrates superior performance in generating physically plausible videos.

As summarized in Table 3.1, PIRA's design offers a uniquely stable, generalizable, and interpretable framework. While VFM-based alignment is sensitive to mismatched latent spaces and correlational black-box teachers, the alternative strategy using a motion-specialized encoder is hampered by its VDM dependency and ill-posed encoder pre-trained to produced intermediate representations for fusion, a task fundamentally different than representation alignment. PIRA is the only approach whose architecture is expressly tailored for representation alignment, pairing a first-principles teacher with a framework that is model-agnostic, efficient, and inherently compatible with the student's latent space.

Attribute	**VFM Alignment**	**Motion-Specialized Encoder**	**PIRA (Ours)**
Teacher Knowledge Source			
Proxy / First-Principle Physics Informed	✗	✓	✓
Interpretable Teacher Signal	✗	✗	✓
Architectural Design			
Inherent Latent Compatibility	✗	✓	✓
Model Agnostic (No Backbone Dependency)	✓	✗	✓
No Specialized Encoder Dependency	✓	✗	✓
High-Quality Targets (No Input Compromises)	✗	✓	✓

Table 3.1: Comparison of representation alignment strategies for physics-aware finetuning. PIRA offers a uniquely stable, generalizable, and interpretable framework by re-purposing the VDM's native VAE as a versatile physics encoder. Symbols: ✓= Advantageous, ✗= Disadvantageous.

Within the PIRA framework, the alignment between the derived teacher targets $\mathbf{y}^{\star}_{\mathbf{physics}}$ and the student's noisy latent features $\mathbf{h_t}$ is performed at a pre-defined intermediate k-th VDM layer. The final component of the PIRA framework is the set of objective functions that carry out the representation alignment. Specifically,

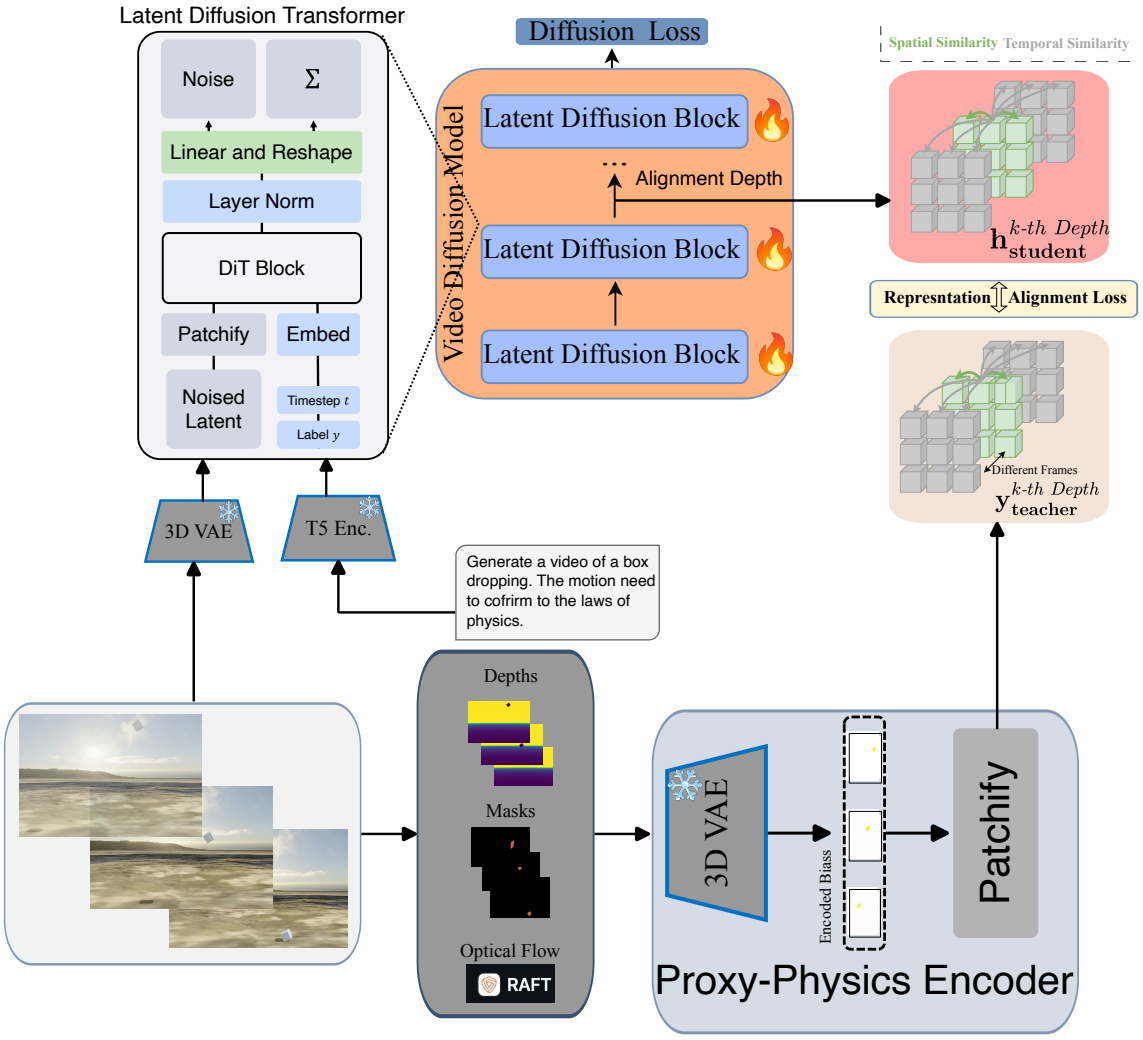


Figure 3.1: The Physics-Informed Representation Alignment (PIRA) architecture. We feed explicit physics signals (e.g., masks, depth, optical flow) through the VDM's *native* 3D-VAE to create inherently compatible teacher representations $\mathbf{y}^{\star}_{\mathbf{physics}}$. These targets are then used with the representation alignment losses to finetune the VDM.

we extend the Vision-Foundation Loss (VF-loss) Yao et al. (2025) from spatial to spatio-temporal space (See Section 3.3), to enforce both cross-frame and intra-frame consistency, thereby mitigating the instability of the "hard alignment" objective, as described in Section 3.1.

3.3 Representation Alignment Objectives

To realize Physics-Informed Representation Alignment, we need a loss objective that can effectively regularize the VDM's hidden states towards our proxy-physics targets ($\mathbf{y}^{\star}_{\text{physics}}$) without inducing the training instability associated with naive feature matching between likely incompatible feature spaces (those of VDMs and our encoded physics representations) that the direct cosine similarity imposes. Drawing inspiration from the robust objectives proposed in (Yao et al., 2025), we extend two complementary alignment losses to the spatio-temporal domain, similarly to the Token Relational Distillation logic introduced in (Zhang et al., 2025b). Both objectives directly address the challenges arising from the hard-alignment of incompatible spaces, providing a robust mechanism for instilling the structure of our physics-informed targets into the VDM.

We first establish a unified notation. Let $\mathbf{h} \in \mathbb{R}^{F\times H\times W\times D'}$ be the student's hidden features from an intermediate VDM layer, where F is the number of frames, $H \times W$ are the spatial dimensions, and D' is the feature dimension. Let $\mathbf{y}^{\star} \in \mathbb{R}^{F\times H\times W\times D}$ be our physics teacher representation. When necessary, a simple

feed-forward projection aligns the student's feature dimension, such that $\mathbf{h}' = \text{Proj}(\mathbf{h}) \in \mathbb{R}^{F\times H\times W\times D}$, as to maximize the guidance from the $\mathbf{y}^\star$.

3.3.1 (MARGINAL) COSINE SIMILARITY LOSS

The Marginal Cosine Similarity loss provides an unbroken, point-to-point alignment signal. It enforces that the student's feature vector at a specific spatio-temporal location should be directionally aligned with the corresponding teacher vector. A linear projection is necessary here to match feature dimensions D' and D.

The loss is composed of two distinct components:

Spatial Point-wise Similarity This measures the direct cosine similarity between student and teacher vectors at identical spatio-temporal locations. The resulting similarity tensor is $u_{\text{spatial}} \in \mathbb{R}^{F\times H\times W}$. For each frame f and spatial location (i, j):

$$u_{\text{spatial}}^{f,i,j} = \frac{\mathbf{y}^{\star,f,i,j} \cdot \mathbf{h}'^{f,i,j}}{\|\mathbf{y}^{\star,f,i,j}\| \cdot \|\mathbf{h}'^{f,i,j}\|} \tag{3.1}$$

Temporal Point-wise Similarity This measures the similarity of features at the same spatial location across different frames, enforcing temporal consistency. The resulting similarity tensor is $u_{\text{temporal}} \in \mathbb{R}^{F\times H\times W\times(F-1)}$. For each spatial location (i, j) and pair of frames (f, f') where $f \neq f'$:

$$u_{\text{temporal}}^{f,f',i,j} = \frac{\mathbf{y}^{\star,f,i,j} \cdot \mathbf{h}'^{f',i,j}}{\|\mathbf{y}^{\star,f,i,j}\| \cdot \|\mathbf{h}'^{f',i,j}\|} \tag{3.2}$$

Total Loss ($\mathcal{L}_{mcos}$) The total loss combines the spatial and temporal components. Specifically, for each component of the total loss, minimize the point-wise similarity distance and subtract by a margin. The ReLU function provides tolerance by only penalizing pairs whose similarity falls below a certain threshold, effectively guiding the alignment towards less similar pairs.

$$\begin{aligned}\mathcal{L}_{mcos} = &\underbrace{\frac{1}{FHW}\sum_{f=1}^{F}\sum_{i=1}^{H}\sum_{j=1}^{W}\text{ReLU}(1 - m_1 - u_{\text{spatial}}^{f,i,j})}_{\text{Spatial Loss}} \\ &+ \underbrace{\frac{1}{F(F-1)HW}\sum_{\substack{f,f'=1\\ f\neq f'}}^{F}\sum_{i=1}^{H}\sum_{j=1}^{W}\text{ReLU}(1 - m_1 - u_{\text{temporal}}^{f,f',i,j})}_{\text{Temporal Loss}}\end{aligned} \tag{3.3}$$

3.3.2 (MARGINAL) DISTANCE MATRIX SIMILARITY LOSS

The Marginal Distance Matrix Similarity loss enforces a more structural alignment objective by bringing the distance matrices of features that correspond to some relative distribution closer. It operates not on the feature vectors themselves, but on the geometry of the feature space, establishing that the matrix of pairwise similarities between all patches in the student's space is structurally similar to the corresponding matrix in the teacher's space. For this formulation, we consider the flattened representations where $S = H \times W$.

Intra-Frame (Spatial) Relational Structure. We first compute the matrix of pairwise cosine similarities between all spatial patches within each frame. This captures the intra-frame geometry. The resulting similarity matrices for the student and teacher, $\mathbf{u}_{\mathbf{h}}^{\text{spatial}}, \mathbf{u}_{\mathbf{y}^\star}^{\text{spatial}} \in \mathbb{R}^{F\times S\times S}$, have elements defined as:

$$u_{h,\text{spatial}}^{f,i,j} = \frac{\mathbf{h}^{f,i} \cdot \mathbf{h}^{f,j}}{\|\mathbf{h}^{f,i}\| \cdot \|\mathbf{h}^{f,j}\|} \quad \text{and} \quad u_{y^\star,\text{spatial}}^{f,i,j} = \frac{\mathbf{y}^{\star,f,i} \cdot \mathbf{y}^{\star,f,j}}{\|\mathbf{y}^{\star,f,i}\| \cdot \|\mathbf{y}^{\star,f,j}\|} \tag{3.4}$$

Inter-Frame (Temporal) Relational Structure. Similarly, we compute the similarity between every patch in a given frame and every patch in every other frame. This captures the complete spatio-temporal geometry. The resulting matrices $\mathbf{u}_{\mathbf{h}}^{\text{temporal}}, \mathbf{u}_{\mathbf{y}^\star}^{\text{temporal}} \in \mathbb{R}^{F\times F\times S\times S}$ have elements defined for $f \neq f'$ as:

$$u_{h,\text{temporal}}^{f,f',i,j} = \frac{\mathbf{h}^{f,i} \cdot \mathbf{h}^{f',j}}{\|\mathbf{h}^{f,i}\| \cdot \|\mathbf{h}^{f',j}\|} \quad \text{and} \quad u_{y^\star,\text{temporal}}^{f,f',i,j} = \frac{\mathbf{y}^{\star,f,i} \cdot \mathbf{y}^{\star,f',j}}{\|\mathbf{y}^{\star,f,i}\| \cdot \|\mathbf{y}^{\star,f',j}\|} \tag{3.5}$$

Because this loss compares scalar similarity values computed within each respective feature space, a trainable projection to align the feature dimensions of $\mathbf{h}$ and $\mathbf{y}^\star$ is not strictly required. We thus treat it as an optional design choice.

The final loss penalizes the absolute difference between the student and teacher similarity matrices, relaxed by a margin m_2. This encourages the relational structures to align without enforcing a strict one-to-one vector correspondence.

$$\mathcal{L}_{mdms} = \underbrace{\frac{1}{FS^2}\sum_{f=1}^{F}\sum_{i,j=1}^{S}\text{ReLU}\left(|u^{f,i,j}_{h,\text{spatial}} - u^{f,i,j}_{y^\star,\text{spatial}}| - m_2\right)}_{\text{Spatial Loss}} + \underbrace{\frac{1}{F(F-1)S^2}\sum_{\substack{f,f'=1\\ f\neq f'}}^{F}\sum_{i,j=1}^{S}\text{ReLU}\left(|u^{f,f',i,j}_{h,\text{temporal}} - u^{f,f',i,j}_{y^\star,\text{temporal}}| - m_2\right)}_{\text{Temporal Loss}} \tag{3.6}$$

3.3.3 Final Training Objective and Integration

We integrate our PIRA objective into a supervised fine-tuning (SFT) process. This fine-tuning is conducted on our purpose-built synthetic dataset of falling dynamics (detailed in Section 5.2). In accordance to previous representation alignment frameworks (Yu et al., 2024; Zhang et al., 2025b), our physics-based transformations from various proxy biases (e.g., depth, masks, and optical flow) are the representation alignment targets $\mathbf{y}^\star$.

The total loss for fine-tuning the VDM is a weighted sum of the standard diffusion loss ($\mathcal{L}_{\text{LDM}}$), defined in Equation 2.1, and one of the proposed representation alignment objectives. The final objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda\mathcal{L}_{\text{PIRA}} \tag{3.7}$$

where $\mathcal{L}_{\text{PIRA}}$ can be either $\mathcal{L}_{mcos}$ (from Equation 3.3) or $\mathcal{L}_{mdms}$ (from Equation 3.6), and λ is a hyperparameter that balances the contribution of the two loss terms.

Chapter 4

Modeling Explicit Physical Reasoning in Video Diffusion Models - MORPHEUS

Evaluating whether a video generative model truly understands physics is a profound challenge. The predominant methods, employing quantitative trajectory matching (Motamed et al., 2025; Agarwal et al., 2025; Li et al., 2025), have a fundamental limitation: they measure fidelity to a *single ground-truth outcome*. A generated video can be penalized for deviating from the ground-truth trajectory even if its own trajectory is perfectly physically plausible under slightly different, unobserved initial conditions. This issue is particularly relevant for I2V models, where single-frame conditioning leaves key physical parameters of the system (e.g., object mass, friction, or precise initial velocity) unobserved. This ambiguity can lead to scenarios where a physically consistent generation is unfairly scored as a failure. A deeper and more immune evaluation of physical reasoning requires moving beyond trajectory matching and comprehending whether a generated video adheres to the underlying, infallible physical laws themselves, regardless of its specific trajectory.

To fill this gap, we designed **Morpheus** benchmark (Zhang et al., 2025a), a framework capable of evaluating physical reasoning in a two-fold fashion: first by directly measuring consistency with governing equations and secondly the conservation of physical invariants. Morpheus overcomes the limitations of ground-truth-dependent metrics by asking the following fundamental questions:

1. Are the observed dynamics consistent with the governing laws of motion?
2. Do the dynamics preserve physical quantities that we know must be conserved?

As illustrated in Figure 4.1, this approach shifts the evaluation from "is it the right trajectory?" to "is it a *physically possible* trajectory?". The core pipeline first extracts an object's trajectory from a video and then analyzes it with two complementary physics-informed scores.

Dynamical Score This score quantifies the trajectory's adherence to the known governing Ordinary Differential Equation (ODE) of the physical system. To compute this, we fit a Physics-Informed Neural Network (PINN) (Raissi et al., 2017a;b) to the extracted 2D trajectory. A PINN is a neural network that receives a timestep i as input and is trained to approximate the trajectory's coordinates $\mathbf{T}_i$. Its loss function, however, contains a crucial physics-based regularization term, L_{physics}, which penalizes deviations from the known equation of motion. For our **case study of a falling item**, the governing equations are $\ddot{x} = 0$ and $\ddot{y} + g = 0$, where g is the gravitational constant. The L_{physics} term explicitly constrains the network's output to conform to these principles. The score is derived from the Normalized Mean Squared Error (NMSE) between the PINN's best-fit trajectory and the observed trajectory. The final score is defined as $1 - \text{NMSE}$, such that a high Dynamical Score (approaching 1) indicates that the observed motion is highly consistent with the physical laws governing the physical phenomenon.

Physical Invariance Score This score provides a fine-grained and robust analysis by measuring the model's ability to preserve specific physical quantities that should be conserved throughout the motion. For a given trajectory, we derive the time series for several known invariants. Since these quantities should be constant, a low variance in their time series indicates high physical consistency. To transform this variance into a standardized score, we first calculate the standard deviation for each invariant's time series (σ_{inv}) and normalize it by its mean (μ_{inv}) to get the relative standard deviation. This value is then mapped to a score S

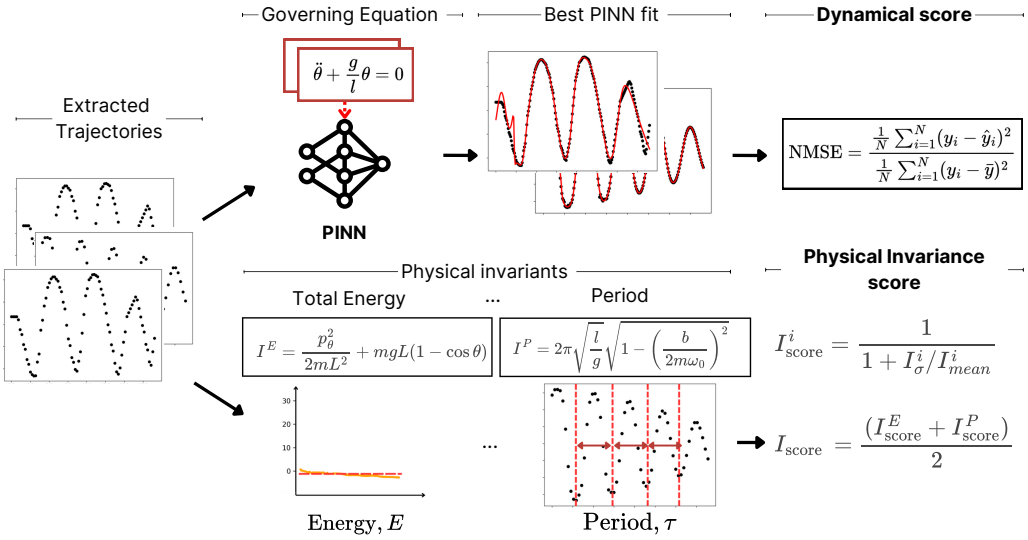


Figure 4.1: Overview of the Morpheus evaluation pipeline. Trajectories obtained from real or generated videos are analyzed using two physics-informed scores. The **Dynamical Score** (top) fits a PINN constrained by the governing ODE to the trajectory. The **Physical Invariance Score** (bottom) computes the variance of conserved quantities like total energy. Figure adjusted from (Zhang et al., 2025a)

in the range $[0, 1]$ using the formula $S = 1/(1+\alpha \cdot (\sigma_{inv}/|\mu_{inv}|))$, where α is a scaling hyperparameter. This ensures that a low relative standard deviation results in a high score, signifying successful conservation. For our **free-falling items** case, and assuming negligible air resistance, we evaluate the following key invariants:

- **Total Energy:** The total energy of the falling item, which is the sum of its kinetic energy $K = \frac{1}{2}m(v_x^2 + v_y^2)$ and potential energy $U = mgy$, must be conserved. While the mass m is unknown, we can still assess the conservation of the energy-to-mass ratio, $E/m = \frac{1}{2}(v_x^2 + v_y^2) + gy$, which should also be constant. We compute the score for this invariant based on the variance of this quantity over the trajectory.
- **Vertical Acceleration:** As the only significant force acting on the object is gravity, its vertical acceleration a_y should be constant and equal to g. We compute a score by measuring the variance of the estimated vertical acceleration over time.
- **Horizontal Momentum:** With no horizontal forces, the horizontal momentum $p_x = mv_x$ is conserved. This implies that the horizontal velocity v_x must remain constant. We compute a score based on the variance of the horizontal velocity.

The final Physical Invariance Score reported in our results is the **average of the individual scores** computed for each of these three invariants. This provides an interpretable, multi-faceted measure of a model's ability to respect fundamental conservation laws. A model might, for instance, perfectly conserve horizontal momentum (high score) but fail to conserve energy (low score), revealing specific and actionable shortcomings in its physical reasoning.

Crucially, this evaluation methodology does not require a ground-truth instance for comparison. Nevertheless, the synthetic test-splits of video we created serve as an empirical upper bound, defining the best possible score a VDM could possibly achieve. The development of this principled evaluation framework was a key motivation for our work; by first establishing a robust way to measure physical plausibility, we were then able to design and validate PIRA. By employing Morpheus, we can decouple our evaluation from specific outcomes and directly quantify the extent to which our finetuning has instilled a genuine, first-principles understanding of physical laws into the VDM.

Chapter 5

Modeling Falling Dynamics with Synthetic Data

This section details our focused approach to modeling a fundamental physical phenomenon as a testbed for our physics-informed representation alignment framework. We first motivate our choice to study the free-fall dynamics of objects, situating it within the broader context of benchmarking physical reasoning in Video Generative Models (VGMs). Subsequently, we describe our methodology for creating a large-scale, controlled synthetic dataset, which is crucial for both effective fine-tuning, which any representation alignment method requires, and rigorous evaluation.

5.1 Modeling Intuitive and Falling Dynamics

The remarkable visual fidelity and realism achieved by VGMs have ignited interest in their potential as scalable world simulators (Brooks et al., 2024; Agarwal et al., 2025). This has stimulated a critical line of research questions related to the extent to which these models truly understand the physical laws that govern our world and whether they can model intuitive physics dynamics. A growing body of work has sought to answer this question by establishing rigorous evaluation benchmarks. These range from VLM-based assessments of "physical commonsense" on complex, real-world interactions (Bansal et al., 2024; 2025; Gu et al., 2025) to quantitative metrics that match generated trajectories against ground-truth data (Li et al., 2025; Agarwal et al., 2025; Motamed et al., 2025). The most principled of these, such as Morpheus (Zhang et al., 2025a), move beyond mere trajectory matching or VQA-bounded score provided by VLM experts (which can have prominent hallucinations) or humans, to evaluate the conservation of physical invariants (e.g., energy, momentum), and fitting the linear ODEs to the underlying equations of motion. A consistent finding across this diverse literature is that even state-of-the-art VGMs frequently fail to model even the most fundamental physical dynamics.

In this work, we limit our evaluation methodology to one such fundamental task: modeling object free-fall under uniform gravity. While seemingly simple, all benchmarks that quantify physical-reasoning have the leading Image-to-Video (I2V) models consistently struggle to generate physically accurate dropping behavior, often producing videos where objects float, move with incorrect acceleration, or violate object permanence (Zhang et al., 2025a; Li et al., 2025). By focusing on this well-defined yet challenging problem, we can conduct controlled experiments to rigorously assess the efficacy of our method in instilling physical knowledge. While we argue PIRA applies to a broader range of Newtonian dynamics, free-fall serves as a perfect, minimal test case to verify our methodology, while staying within our budget limits.

5.2 A Principled Case for Synthetic Data

To effectively instill knowledge of a specific physical law, a clean, controlled, and large-scale data source is vital. We contend that for this purpose, a synthetic dataset generated from a high-fidelity simulator is not only scalable but a superior foundation for several key reasons.

Fine-grained Control, Rich Annotation-Levels Simulators provide complete control over the physical process. We can precisely vary initial conditions like object height, mass, and scene depth, and in turn, obtain

perfect, denoised ground-truth annotations for our teacher bias signals (e.g., segmentation masks, depth maps, and many more if needed), allowing us to generate a vast and diverse distribution of scenarios governed by the same underlying dynamics. More importantly, this paradigm provides access to perfect, noise-free, multi-modal ground-truth annotations at virtually no cost. Our direct mapping to physical state variables (evolving over time, like per-frame segmentation masks, dense depth maps) can be extracted directly by the simulation state. Such a level of control and annotation fidelity cannot be reached with internet-scale, in-the-wild internet data. This is highlighted by the recent, large-scale efforts of works like WISA (Wang et al., 2025) framework addressing the need for clean, physics-centric data by undertaking a laborious process of manual video collection involving a lot of human resources, followed by LLM-intensive annotation. While valuable, such an approach is hampered by the massive human effort and the subjective nature of curation, limiting its scalability. At the other spectrum NVIDIA's vision for generalist pre-trained World Foundation models (Agarwal et al., 2025) demonstrate that a usable signal from a massive collection of both public and proprietary datasets with diverse categories (e.g. Spatial awareness and navigation, nature dynamics, driving, Hand motion and object manipulation and etc.) requires a sophisticated, multi-stage filtering pipeline (including curation, motion and quality filtering, duplications, annotation). Even after multiple rounds of such processing, the resulting data do not guarantee explicit ground-truth physical parameters or mapping to physical state variables. The latter is essential for the targeted, first-principles knowledge we go after. Our synthetic approach elegantly sidesteps all these challenges.

Alignment with VDM specific design choices during pre-training A significant practical challenge in fine-tuning open-source VGMs is their sensitivity to the specific data configurations used during pre-training. The released pre-trained model weights, along with the capability of generation models, are often fixed to a particular resolution, aspect ratio, and frame rate (FPS) due to design choice and training paradigms applied in any preceding pre-training round(s). Attempting to fine-tune them on data that deviates from these specifications can lead to training instability, degraded performance, or even pure noise generation. Real-world videos, with their variable formats, require extensive and often lossy pre-processing (like applying computer vision techniques like interpolation and cropping) to conform to these strict constraints, without any guarantee on the final Signal-to-Noise ratio. In contrast, a simulator allows us to generate a large-scale, flexible dataset that perfectly and natively matches the exact specifications of any target VDM backbone. This ensures a stable and maximally effective Supervised Fine-Tuning (SFT) process, allowing the model to focus solely on learning the new physical knowledge without being confounded by domain shifts in data format.

Facilitating Generalization Tests A primary goal of our work is to ensure the VDM does not merely overfit to the fine-tuning distribution but truly learns the underlying physical principle. A synthetic environment is uniquely suited for this purpose, as it enables the creation of carefully controlled test splits. By reserving a subset of 3D assets and background environments exclusively for evaluation, we can trivially construct distinct *in-distribution* (ID) test sets, featuring objects and scenes seen during training, and *out-of-distribution* (OOD) test sets with entirely novel objects and backgrounds. This allows us to rigorously and quantitatively measure the model's ability to generalize the learned physical law to unseen scenarios, providing a transparent and objective assessment of its progress towards true physical reasoning.

For our large-scale SFT dataset creation, we used the Kubric simulation framework (Greff et al., 2022) [1]. In total, we create 10,000 synthetic samples of falling items following the normal Earth's gravitational pull[2]. Each sample is configured to match the specifications of our VDM backbone, CogVideo-X (Yang et al., 2024), with a duration of 49 frames at 8 FPS.

We populated our samples with a wide variety of objects imported from photorealistic 3D models from the Google Scanned Objects (GSO) dataset (Downs et al., 2022) (See Figure 5.2), to promote sample diversity and prevent overfitting to specific appearances and or motions. The distribution of objects that our SFT dataset consists of is detailed in Figure 5.1. The objects were rendered against a set of 426 unique HDRI backgrounds. Throughout each video, the camera remains stationary and oriented parallel to the ground plane, ensuring that the learning signal is focused exclusively on object dynamics. Finally, to fairly assess the generalization capabilities of our method, we curated two distinct test splits of 64 videos each: an *in-distribution* (ID) set containing objects and backgrounds seen during fine-tuning, and an *out-of-distribution* (OOD) set featuring unobserved assets.

[1] Kubric leverages the PyBullet (Coumans & Bai, 2016–2021) physics engine and the Blender (Community, 2018) renderer,

[2] $9.81 m/s^2$

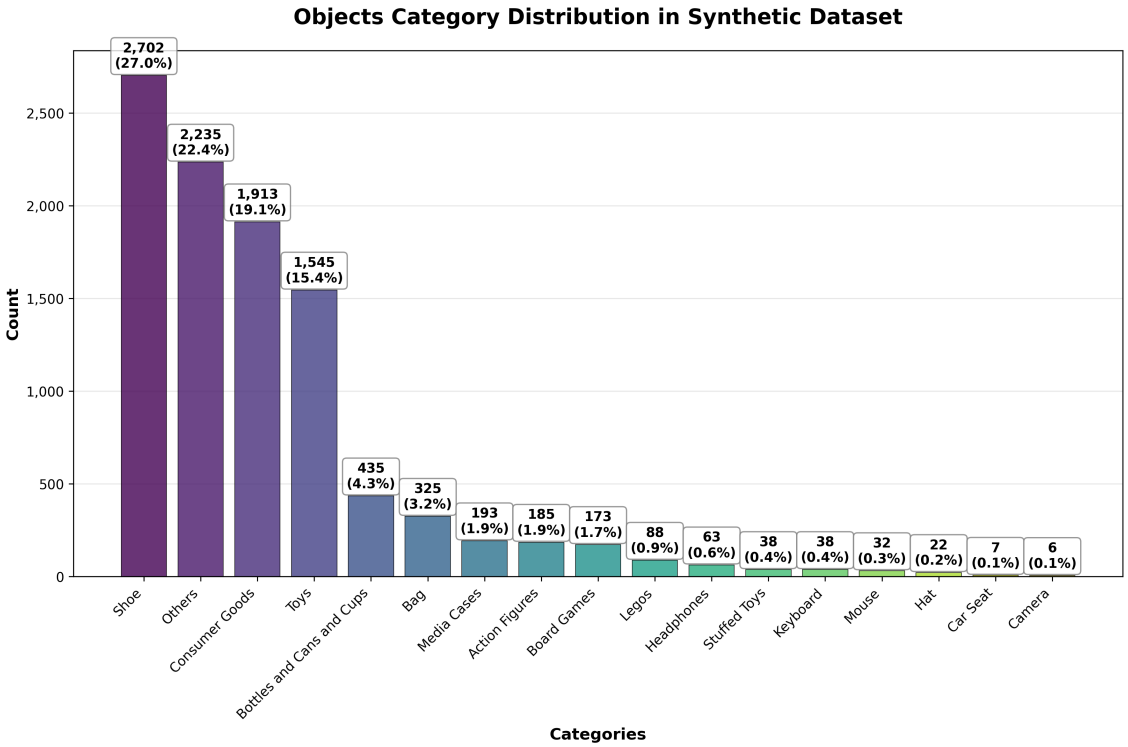


Figure 5.1: Statistics of the synthetic dataset - Categories/Types of dropping objects created for SFT split.

Figure 5.2: Examples of object sampled from the GSO dataset Downs et al. (2022) that are featured in our SFT samples.

Chapter 6

Experimental Framework & Results

In this chapter we comprehensively cover our experimental framework designed to validate the effectiveness of our proposed PIRA framework. We start with outlining the implementation details associated with our physics-informed representation alignment strategy, including the model backbone, dataset configurations, and the specific setups for our PIRA variants and baselines. We then proceed with a qualitative evaluation, showcasing visual comparisons that provide an intuitive understanding of PIRA's impact on generating physically plausible videos. Next, we delve into our two-folded quantitative analysis, first assessing performance through implicit state variable matching metrics and then through our explicit, physics-informed *Morpheus* benchmark(Chapter 4), providing a systematic comparison between our proposed methods and relevant baselines. Finally, we conclude with a series of in-depth ablation studies to analyze the sensitivity and contribution of PIRA's key components and hyperparameters.

6.1 Implementation Details

Model Setup Across all our experiments, we use the open-source **CogVideoX-5B-I2V** (Yang et al., 2024) as our VDM backbone. This powerful, expert transformer-based Image-to-Video model serves as an ideal testbed for our representation alignment finetuning. Our experimental setup consists of two main methodological branches:

- **Specialized Motion Encoder Variant:** For this approach, we use the pre-trained Trajectory Extractor (TE) and its associated motion-specific 3D-VAE, as introduced in TORA (Zhang et al., 2025c). The teacher signal is derived by feeding ground-truth optical flow maps into this TE. We refer the avid reader to Appendix A.1 for a detailed walkthrough of design of this physics-transformation based variant. Note that in our physical-reasoning quantification section (Section 6.3) we use the terms *TORA-based* and *Specialized Motion Encoder* interchangeably to refer to this variant.
- **PIRA (Ours):** Our proposed PIRA framework, employs the VDMs (CogVideoX-5B-I2V) own **native 3D-VAE** as a capable encoder for encoding proxy-physics centric bias. We test three distinct physics-proxy signals by feeding them directly into this native VAE to create the teacher representations:
  1. **Masks:** Ground-truth segmentation masks rendered directly from the simulator.
  2. **Depth:** Ground-truth depth maps rendered directly from the simulator.
  3. **Optical Flow:** Optical flow maps generated by applying RAFT (Teed & Deng, 2020) to the rendered synthetic videos.

Dataset and Training Procedure. All models are finetuned on our synthetic dataset of 10,000 videos depicting falling objects, as detailed in Section 5.2. To ensure computational efficiency and preserve the model's pre-trained knowledge, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) for all finetuning experiments. We apply LoRA to the attention blocks of the VDM's DiT with a rank of $r = 128$ and an alpha of $\alpha = 64$. Each model variant is trained for a total of two epochs. For fair comparison, all baseline models (e.g., finetuned CogVideoX) and our proposed methods (TORA-based and all PIRA variants) are trained for the exact same number of steps using identical LoRA configurations. A single finetuning run

takes approximately 20 hours on 4 H100-SXM5-94GB GPUs, using a micro-batch size of 2 per device and gradient accumulation after every single gradient step, achieving an effective batch size of 8.

Representation Alignment Configuration. The core of our method involves injecting the representation alignment loss at an intermediate layer of the VDM's denoising transformer. Following the findings of prior work, such as VideoREPA (Zhang et al., 2025b), which identified that mid-to-late layers are optimal for distilling high-level knowledge, we consistently carry out the representation alignment with our physics-informed targets to the output of the **18th transformer block** of the VDM backbone (*CogVideoX-5B-I2V*).

6.2 Qualitative Evaluation

We begin our experimental analysis with a qualitative evaluation to provide an intuitive understanding of our method's impact before delving into quantifying the physical reasoning. By showcasing the visual results first, we aim to highlight the prominent improvements PIRA brings to a pre-trained Video Diffusion Model in terms of physical plausibility. We present comparisons on two distinct test splits, which serve to evaluate PIRA's ability to both master the learned dynamics and generalize them to novel scenarios.

To visualize object motion, we employ a mask-based motion overlay technique[1]. Each image presented in this section is a static composite created by extracting the falling object from multiple, adaptively selected video frames and overlaying these snapshots onto the initial frame. To convey the temporal sequence, the extracted objects are blended with progressively increasing opacity, from more transparent (earlier in the fall) to fully opaque (later in the fall). The purpose of this approach is to produce a single, clear image that captures the object's complete trajectory and dynamics. With this visualization technique we compare our **PIRA** method against the original **CogVideoX-5B-I2V** (Yang et al., 2024) baseline and the physically accurate **Ground Truth** videos created in the simulator (process described in Section 5.2).

Generalization to Unseen Scenarios (OOD Split) First, we test PIRA's generalization ability. Figure 6.1 illustrates trajectories on the out-of-distribution (OOD) test split, which contains different objects and backgrounds than the ones featured during the SFT process.

The ground-truth videos (Figure 6.1a) serve as our reference, consistently depicting objects undergoing smooth, vertical acceleration due to gravity. Our **PIRA** framework (Figure 6.1b) demonstrates unprecedented ability to generalize physical principles. Its generated trajectories closely emulate the falling dynamics of the ground truth samples, exhibiting correct vertical motion and preserving object permanence, even in these novel scenarios. The latter underscores PIRA's success in instilling truly generalizable understanding of free-fall dynamics.

In stark contrast, the original **CogVideoX baseline** (Figure 6.1c), fundamentally fails to model this simple physical phenomenon. Its generations are plagued by severe artifacts and disregard physical commonsense, including entirely motionless generations, physically impossible trajectories (e.g., pure horizontal movement), and a catastrophic failure of object permanence where single objects dissolve into multiple distorted copies.

Performance on Seen Scenarios (ID Split) Next, we assess PIRA's performance on the in-distribution (ID) test split, containing objects and backgrounds that were part of the SFT process. As shown in Figure 6.2, our **PIRA** method (Figure 6.2b) consistently produces smooth, accelerating trajectories that are nearly indistinguishable from the ground truth (Figure 6.2a).

The **CogVideoX baseline's** performance on this split (Figure 6.2c) further highlights its core limitations. Keep in mind that the term *In-Distribution* does not have a particular meaning in this case as the data are still unprecedented for the baseline VDM, which lacks physical reasoning. While many of the failure modes experienced in the OOD's generation repeat, new and equally severe artifacts also emerge, such as severe object deformation and smearing and spontaneous disappearance and teleportation. This confirms that the baseline's inability to model physics is fundamental and not dependent on the specific assets presented.

This dramatic qualitative gap—persistent across both test splits—serves as evidence of our framework's effectiveness. The baseline's inability to model even the simplest gravitational dynamics highlights the necessity of an approach which successfully guides VDMs to generate videos that are both visually recognizable and physically sound. Our qualitative comparisons motivate the subsequent quantitative sections, where we will rigorously measure the extent to which PIRA closes this physics gap.

[1]The masks of the generated video were obtained with the SAM-2 video predictor. This process is already comprehensively described in Section 6.3.1

(a) Ground Truth (b) PIRA (Ours) (c) Baseline CogVideoX

Figure 6.1: Qualitative comparison on the **OOD test split**. Each image visualizes motion by overlaying frames. **(a)** Ground-truth videos show correct vertical free-fall. **(b)** Our PIRA method successfully *generalizes* to generate physically plausible trajectories for unseen objects. **(c)** The original CogVideoX baseline exhibits a complete failure to model gravity.

6.3 Quantifying Physical Plausibility

Following our qualitative showcase, this section transitions to the core quantitative validation of our work. Measuring the physical realism of generated video is a non-trivial task. While early benchmarks relied on subjective VLM-based scoring (Bansal et al., 2024; 2025; Gu et al., 2025), the field has moved towards more objective-driven evaluations. To provide a comprehensive and principled assessment, we structure our analysis into two distinct paradigms.

In the first Sub-Section 6.3.1, we first define and then evaluate the performance PIRA and the presented baselines acquire, based on *implicit physical plausibility*. This approach uses state variable matching (e.g., comparing object trajectories and shapes) to measure how closely a generated video's outcome matches a ground-truth reference. Subsequently, in Sub-Section 6.3.2, we evaluate our models using our *explicit physical plausibility* benchmark, *Morpheus*. As introduced in Chapter 4, Morpheus moves beyond specific outcome-matching to directly quantify a model's adherence to governing physical laws and conservation

(a) Ground Truth (b) PIRA (Ours) (c) Baseline CogVideoX

Figure 6.2: Qualitative comparison on the **ID test split**. **(a)** Ground truth. **(b)** PIRA generates physically correct motion for objects seen during fine-tuning. **(c)** The baseline continues to fail, exhibiting artifacts like object **deformation** and **smearing**, confirming its fundamental inability to reason about physics.

principles. By presenting results from both implicit and explicit frameworks, we demonstrate PIRA's superior performance but also uncover deeper insights into the nature of the physical understanding it distills.

As a reminder, throughout this whole section we use the terms *TORA-based* and *Specialized Motion Encoder* interchangeably to refer to the same variant, presented in Appendix A.1.

6.3.1 Implicit Physical Plausibility Quantification via State Variable Matching

A robust way to evaluate a model's physical understanding is to compare the physical state variables associated with the generated video against a ground-truth instance. For a visual scene, these states correspond to object trajectories, shaped and spatial extents over time. Recent benchmarks such as PisaBench (Li et al., 2025) have championed this approach, leveraging the simulation data, which serve as the ground-truth reference, to perform direct, quantitative comparisons. Other works, while not always leveraging physics simulators, have also proposed similar pixel-level metrics associated with some type of trajectory matching for their evaluations (Motamed et al., 2025; Agarwal et al., 2025).

Given that PIRA is finetuned on synthetic data with access to perfect ground-truth annotations, this evaluation paradigm is exceptionally well-suited for our framework. It allows us to move beyond mere perceptual realism and measure VDM's adherence to the physical dynamics. In our strategy, we group these metrics by their semantics and conceptual focus, resulting in three distinct families.

PisaBench-style Trajectory and Shape Metrics To assess the model's grasp of an object's path and form, we adapt metrics from PisaBench (Li et al., 2025), which provide robust measures of geometric deviation. Using the 2D centroid of the object's segmentation mask in each frame, we compute:

- **Trajectory L2 Distance:** The average Euclidean distance between the generated and ground-truth centroids at each frame. This provides a direct measure of overall trajectory error and the model's ability to reproduce the correct equations of motion.
- **Chamfer Distance (CD):** While L2 distance is sensitive to point-wise errors, Chamfer Distance assesses the fidelity of the overall path shape by measuring the average closest-point distance between the two trajectory point sets, making it more robust to minor temporal misalignment.

Physics-IQ-style Object Permanence Metrics Beyond the trajectory, a physically plausible video must maintain the integrity of the object itself. Inspired by metrics used in works like Physics-IQ (Motamed et al., 2025), we use several variants of Intersection over Union (IoU) to evaluate object permanence, shape consistency, and motion accuracy.

- **Spatiotemporal IoU (ST-IoU):** Computes the IoU of the object's segmentation mask frame-by-frame and averages over time. A high ST-IoU demands that the generated object is at the right place, at the right time, with the right shape.
- **Spatial IoU (S-IoU):** Collapses the temporal dimension by taking the union of all masks over time before computing IoU. This metric assesses *where* action happened, relaxing the constraint on *when* it happened.
- **Weighted Spatial IoU (W-S-IoU):** Similar to S-IoU, but it weights the spatial map by the frequency of pixel activation over time. This metric assesses not only *where* action happened but also *how much* action happened in each location.

PIRA's Occlusion-Robust Trajectory Metrics To complement geometric fidelity metrics like trajectory L2 Distance and Chamfer Distance from PisaBench (Li et al., 2025), we introduce two auxiliary Mean Squared Error (MSE) based metrics that also operate on object centroids. A key limitation we observed, particularly in the available implementations of prior benchmarks, is the lack of robust handling for tracking failures and object occlusions. In some generated videos, especially in generic pre-trained VDMs, objects may temporarily vanish for several frames before reappearing or being occluded. To address this, we designed our MSE metrics to be robust against such missing data while providing consistent and interpretable scores across diverse conditions.

- **Variance-Normalized MSE (MSE-var):** This metric is designed to provide a score that is invariant to the *scale of motion*. A simple pixel-based MSE is problematic because a 10-pixel error for a nearly static object is a severe failure, while the same error for an object traversing hundreds of pixels might be negligible. Our MSE-var solves this by normalizing the squared error by the total variance of the ground-truth trajectory, computed from the trace of its 2D covariance matrix. This ensures that the error is always relative to the amount of motion in the scene, allowing for fair comparisons between different dynamic scenarios. Furthermore, it incorporates a gradual penalty for occluded frames, ensuring that models are appropriately penalized for losing track of an object.
- **Spatial-Normalized MSE (MSE-spat):** This metric is designed to provide a score that is invariant to *image resolution and aspect ratio*. A fixed pixel-level error metric has a different meaning across videos with different resolution and aspect ratios. If implemented naively the comparisons across different models or datasets become unreliable. To account for arbitrarty resolution and aspect ratio, our MSE-spat. normalizes the squared error by the squared diagonal of the video frame. This makes the metric dimensionless and directly comparable across different evaluation settings, regardless of their native output resolution. Like MSE-var, it also features a robust penalty mechanism to account for occlusions, ensuring that object disappearance is appropriately reflected in the final score.

A significant advantage of this evaluation approach is its practical convenience. For both our synthetic and any real-world ground-truth videos, we only need segmentation masks to derive all three metrics. Since we

operate in the Image-to-Video (I2V) setting, the initial frame is identical for both the generated and ground-truth videos. We can therefore use a state-of-the-art tracker like SAM-2 (Ravi et al., 2024) to conveniently propagate the initial-frame mask throughout the ground-truth video, automatically generating the necessary annotations for all metrics. Together, these metrics provide a comprehensive and interpretable evaluation framework. The trajectory-based metrics assess the model's grasp of dynamics, while the object-level IoU assesses its understanding of object permanence and rigidity. This allows us to systematically analyze how different PIRA configurations contribute to the final video's physical plausibility, yet they remain proxies for the underlying physical laws themselves.

Quantitative Analysis of Implicit Physical Plausibility

We now present an in-depth analysis of our results on implicit physical plausibility quantification, evaluating each proposed method against the metrics detailed in Subsection 6.3.1. Our results are summarized in Tables 6.1, 6.2, and 6.3, which respectively present performance on PisaBench-style metrics, our occlusion-robust metrics, and Physics-IQ-style metrics. Each table reports the best-performing configuration for every methodological approach, abstracting away the hyperparameters of our PIRA framework (e.g, choice of representation alignment loss, marginal values choices in losses, use of MLP projection when $\mathcal{L}_{mdms}$ applied) to enable a fair, high-level comparison.

The *Oracle* entry in each table establishes an empirical upper bound by using SAM-2 predictions on the ground-truth videos instead of using ground-truth masks (and assuming a perfect score always), thereby accounting for any error inherent in the tracking pipeline itself, which is used for evaluating synthetic videos as well. All results are reported in the two synthetic test splits, namely the ID and OOD, respectively, which we explicitly created to quantify the physical commonsense of the modes. Subsection 5.2 describes the creation of these two test-splits.

Baselines Demonstrate the Need for Principled Regularization The performance of our baseline models is revealing. The original, pre-trained CogVideoX model performs poorly across all metrics, confirming that large-scale pre-training on in-the-wild data is insufficient for learning intuitive physics, even for a fundamental task like object freefall. While a standard SFT approach significantly improves performance. For example, reducing the L2 error by nearly half on the OOD split (from 0.052 to 0.021 in Table 6.1), a significant gap to the oracle still remains. This demonstrates that while finetuning, on the synthetic data we created, is highly beneficial, it is not the way forward; a more principled approach is required to distill the underlying physical dynamics into the model effectively.

PIRA Outperforms All Baselines and the TORA-based Approach Our PIRA framework, in all its variants, consistently outperforms both the baselines and the (ill-posed) TORA-based implementation across all metric families. This is one of our central findings. For instance, in the PisaBench-style evaluation (Table 6.1), the Mask-based PIRA variant achieves a remarkable L2 error of 0.012 on the ID split, more than halving the error of the SFT baseline (0.028) while outperforming the TORA-based approach (0.020) by quite some margin. This superior performance trend repeats for the occlusion-robust metrics (Table 6.2) and for object permanence metrics (Table 6.3), where the PIRA approach consistently achieves the highest IoU scores. These results strongly validate our core hypothesis: regularizing a VDM's internal states towards a proxy-physics teacher, encoded with the model's native VAE, is a reasonably effective method for instilling physical plausibility.

Analysis of PIRA Variants A deeper analysis of the PIRA family itself provides several key insights. Across the three proxy-physics signals, we observe highly competitive performance, with the Mask-based variant achieving the best overall results. The strong performance of all three variants (Masks, Depth, and Optical Flow) verifies our hypothesis that the VDM's native 3D-VAE is a versatile feature encoder. It demonstrates that the VAE, despite being pre-trained on high-frequency pixel data, is highly capable of encoding more structured, lower-frequency signals into meaningful and effective representations for alignment. Effectively, the latter alleviates the need for a specialized, motion-specific encoder as in the TORA-based approach, underscoring the modularity and robustness of our framework. While the Mask-based variant excels at ID evaluation, it's worth noting that the choice of proxy is not the only factor; the optimization landscape of the alignment loss itself plays a significant role, and it's plausible that one bias objective is simply easier to optimize than others.

Generalization to Out-of-Distribution Scenarios. The true test of whether a model has learned a physical principle, rather than merely memorizing the training data, is its performance on out-of-distribution (OOD) synthetic test split. Our results in the OOD columns of the tables are particularly illuminating. We observe

that the performance hierarchy shifts. For example, the depth-based PIRA variant emerges as a top performer, achieving the best OOD ST-IoU score (0.524) and S-IoU score (0.688) in Table 6.3. Similarly, the Mask-based PIRA, while still the overall best, shares its top L2 rank with the Depth-based PIRA on OOD data (both achieving 0.013 in Table 6.1). This suggests that while object integrity (Masks) may be a more direct and easier signal to learn for in-domain data, the geometric priors distilled from depth maps may generalize more robustly to unseen objects and backgrounds. The strong OOD performance of all PIRA variants, consistently outperforming the SFT baseline by a large margin, confirms that our physics-informed alignment is not merely overfitting but is successfully instilling a generalizable understanding of free-fall dynamics into the model.

		ID		OOD	
Method		**L2 (↓)**	**CD (↓)**	**L2 (↓)**	**CD (↓)**
Oracle	SAM-2 GT	0.001	0.000	0.000	0.000
Baselines	CogVideoX-5B-I2V	0.048	0.123	0.052	0.130
	SFT CogVideoX-5B-I2V	0.028	0.070	0.021	0.046
TORA based	Specialized Optical-flow Bias	0.020	0.044	0.019	0.036
PIRA	Depth-based Physics	0.015	0.031	0.013	0.022
	Mask-based Physics	0.012	0.020	0.013	0.021
	Optical Flow-based Physics	0.013	0.022	0.014	0.024

Table 6.1: PisaBench-style evaluation focusing on trajectory and shape fidelity. We highlight the best and second-best values for each metric across all competing methods. The Oracle row represents the ground truth and is excluded from the comparison. (↓) Lower is better.

Method		ID			OOD		
		L2(↓)	**MSE (var)(↓)**	**MSE (spat)(↓)**	**L2(↓)**	**MSE (var)(↓)**	**MSE (spat)(↓)**
Oracle	SAM-2 GT	0.001	0.000	0.000	0.000	0.000	0.000
Baselines	CogVideoX-5B-I2V	0.126	0.637	0.451	0.096	0.697	0.455
	SFT CogVideoX-5B-I2V	0.033	0.304	0.244	0.027	0.214	0.197
TORA based	Specialized Optical-flow Bias	0.020	0.191	0.156	0.019	0.186	0.160
PIRA	Depth-based Physics	0.016	0.114	0.110	0.013	0.097	0.094
	Mask-based Physics	0.012	0.072	0.078	0.013	0.092	0.088
	Optical Flow-based Physics	0.013	0.083	0.083	0.014	0.101	0.103

Table 6.2: Evaluation with occlusion-robust trajectory metrics. We highlight the best and second-best values for each metric across all competing methods. The Oracle row represents the ground truth and is excluded from the comparison. (↓) Lower is better.

Method		ID			OOD		
		ST-IoU (↑)	**S-IoU (↑)**	**W-S-IoU (↑)**	**ST-IoU (↑)**	**S-IoU (↑)**	**W-S-IoU (↑)**
Oracle	SAM-2 GT	0.911	0.921	0.911	0.931	0.947	0.931
Baselines	CogVideoX-5B-I2V	0.235	0.403	0.314	0.235	0.398	0.308
	SFT CogVideoX-5B-I2V	0.391	0.594	0.546	0.432	0.651	0.594
TORA based	Specialized Optical-flow Bias	0.460	0.616	0.577	0.473	0.645	0.604
PIRA	Depth-based Physics	0.493	0.629	0.596	0.524	0.688	0.640
	Mask-based Physics	0.525	0.655	0.625	0.517	0.684	0.646
	Optical Flow-based Physics	0.523	0.643	0.612	0.494	0.674	0.633

Table 6.3: Physics-IQ-style evaluation for object permanence and shape consistency. We highlight the best and second-best values for each metric across all competing methods. The Oracle row represents the ground truth and is excluded from the comparison. (↑) Higher is better.

6.3.2 Explicit Physical Plausibility Quantification

Here we demystify the main takeaways of our benchmark *Morpheus*, focusing on the more informative, explicit physics-informed scores. As detailed in Section 4, the proposed metrics provide a principled evaluation of physical reasoning by directly measuring a model's adherence to governing ODEs (Dynamical Score) and its preservation of physical invariants (Physical Invariance Score). The results for all our resulting physics-informed variants and baselines are presented in Table 6.4. For the PIRA approach, the presented entries correspond to the best-performing configuration for each physics-grounded proxy, abstracting away framework hyperparameters (e.g., choice of representation alignment loss, marginal values choices in losses, use of MLP projection in Distance Matrix Similarity) to enable a fair, high-level comparison.

The *Oracle* entry in the table establishes an empirical upper bound by using SAM-2 predictions on the ground-truth videos, thereby accounting for any error inherent in the tracking pipeline itself. All results are reported on the in-distribution (ID) and out-of-distribution (OOD) synthetic test splits, which we explicitly created to quantify the physical commonsense of the models (see Section 5.2).

Baselines: From Visual Patterns to Physical Principles The performance of the baseline models is highly revealing. The original, pre-trained CogVideoX model achieves low scores across the board, confirming that a model trained on diverse, in-the-wild data does not inherently develop a robust understanding of physical laws. Finetuning the VDM backbone, while keeping the architecture and training dynamics unchanged, on our synthetic dataset yields a significant improvement, especially in the Dynamical Score. However, a significant performance gap remains to the oracle's upper-bound, underlining the demand for a more principled approach to instill physical commonsense into VDMs.

PIRA's Superiority in Learning Physical Laws Our PIRA framework, across all its variants, consistently outperforms both the SFT baseline and the TORA-based approach. The PIRA methods achieve the highest scores across all three metric types (Dynamical, Physical Invariance, and Combined), indicating that their generations are most consistent with the true governing equations of falling dynamics. This is evident in both ID and OOD splits. On the ID split, the Mask-based PIRA variant achieves a near-perfect Dynamical Score (0.995), suggesting an almost complete internalization of the governing equations of motion. While the TORA-based method performs decently, its consistently lower scores support our hypothesis that its teacher signal, being optimized for a different objective (motion fusion), is a less effective teacher for instilling fundamental physical laws.

Analysis of PIRA Variants and Generalization Within the PIRA family, we observe highly competitive performance, with all three proxy-physics signals proving to be effective teachers. On the ID split, the Mask-based variant is the clear top performer, achieving the highest scores across all three metrics and nearly matching the oracle's Dynamical Score. This indicates that for in-domain scenarios, aligning to object shape and permanence provides an exceptionally powerful and well-balanced learning signal.

The true test of generalization, however, lies in the out-of-distribution (OOD) results. Here, the performance hierarchy becomes more nuanced, highlighting the distinct inductive biases of each physics proxy. No single variant dominates all metrics; instead, we observe a fascinating specialization:

- The **Depth-based** variant excels at modeling the governing laws of motion, achieving the highest **Dynamical Score**.
- The **Mask-based** proxy, emphasizing object integrity, is most effective at preserving physical invariants, securing the top **Physical Invariance Score**.
- The **Optical Flow-based** variant, which directly encodes motion, strikes the best overall balance, yielding the highest **Combined Score**.

This multifaceted outcome suggests that different proxies instill different, complementary aspects of physical understanding. The consistent, state-of-the-art performance of all PIRA variants on the OOD data strongly indicates that our framework is not merely overfitting but is successfully instilling a generalizable and robust understanding of physical dynamics.

A Deeper Look: The Missing Pattern in Implicit Metrics The explicit, disentangled scores of Morpheus allow us to uncover a pattern that is obscured by the implicit state variable matching metrics. While the TORA-based approach performs competitively on implicit metrics like L2 distance (Table 6.1), Morpheus scores in Table 6.4 reveal a critical weakness: it achieves a relatively strong Physical Invariance Score but a comparatively weaker Dynamical Score. This indicates that a model can learn to generate videos where

certain quantities (e.g., *energy, acceleration*) are roughly conserved, without having learned the correct underlying equation of motion that dictates the object's trajectory. Implicit metrics, which conflate these two aspects, would fail to capture this crucial nuance. Our explicit evaluation, by contrast, clearly demonstrates that the PIRA approach is not only better at preserving invariants but is fundamentally superior at capturing the correct governing dynamics, marking a deeper level of physical understanding.

In conclusion, while implicit metrics demonstrate PIRA's superior ability to match ground-truth outcomes, it is the explicit, physics-informed scores from Morpheus that reveal the deeper reasoning: PIRA is uniquely effective at teaching the VDM to adhere to the fundamental laws of physics that govern a physical phenomenon, while simultaneously respecting the quantities that should remain constant throughout its evolution.

		ID			OOD		
Method		**Dyn. (↑)**	**Phys. (↑)**	**Comb. (↑)**	**Dyn. (↑)**	**Phys. (↑)**	**Comb. (↑)**
Oracle	SAM-2 GT	0.998	0.944	0.971	0.959	0.953	0.956
Baselines	CogVideoX-5B-I2V	0.392	0.638	0.515	0.400	0.743	0.572
	SFT CogVideoX-5B-I2V	0.672	0.782	0.727	0.636	0.822	0.729
TORA based	Specialized Optical-flow Bias	0.897	0.901	0.899	0.889	0.911	0.900
PIRA	Depth-based Physics	0.930	0.912	0.921	0.928	0.912	0.920
	Mask-based Physics	0.995	0.939	0.967	0.902	0.939	0.921
	Optical Flow-based Physics	0.979	0.926	0.953	0.917	0.938	0.928

Table 6.4: **Morpheus Scores - Quantitative Analysis of Explicit Physical Plausibility**. We highlight the best and second-best values for each metric across all competing methods. The Oracle row represents the ground truth and is excluded from the comparison. (↑) Higher is better. Column headers are abbreviated: Dyn. (Dynamical Score), Phys. (Physical Invariance Score), and Comb.(Combined Score $= \frac{1}{2}$Dyn. $+ \frac{1}{2}$Phys.).

6.4 Ablation Studies

To better understand the key components and hyperparameter sensitivities of our PIRA framework, we conducted a series of ablation studies based on the comprehensive experimental results presented in Appendix A.2. We explored a vast hyperparameter space for each physics proxy, varying the choice of representation alignment objective (*Marginal Cosine Similarity* vs. *Marginal Distance Matrix Similarity*), the use of a projection MLP for aligning the dimensionality of diffusion transformer hidden states with the transformer physics-informed representation alignment targets, and the loss marginals (As covered in Chapter3.3). Here, we analyze the impact of these crucial design choices on both implicit and explicit measures of physical reasoning.

Table A.5 illustrates the physics-informed scores from our *Morpheus* benchmark across the various design choices we explored. For each physics proxy, the table highlights the specific hyperparameters and design choices (introduced in Chapter 3.3) that yielded the best-performing model, providing a clear summary of our optimization process.

Point-to-Point vs. Relational Alignment A central design choice in our framework is the type of alignment loss used to distill physical knowledge. Our comprehensive results, detailed in Appendix A.2, reveal a nuanced relationship between the loss function and the teacher signal.

Focusing on our explicit *Morpheus* scores in Table A.5, a clear pattern emerges. For the **Mask-based proxy**, the *point-to-point alignment*(Equation 3.3) is unequivocally superior, delivering the best performance in both ID and OOD splits. For the **Depth and Optical Flow proxies**, however, the *relational*

alignment(Equation 3.6)proves to be the better choice, especially for out-of-distribution generalization. For instance, the Depth-based PIRA with $\mathcal{L}_{mdms}$ (specifically, MDMS No MLP, $m = 0.0$) achieves the top OOD Dynamical score, while the Optical Flow variant with $\mathcal{L}_{mdms}$ (MDMS No MLP, m=0.1) secures the best overall OOD Combined Score.

Interestingly, this pattern is not perfectly mirrored in our implicit, trajectory-matching metrics (Table A.1). While the Mask and Optical Flow proxies show consistent representation alignment objective preferences across both evaluation types, the Depth proxy performs best on implicit metrics with the point-to-point (Equation!3.3) loss (OOD L2 of 0.013). This inconsistency highlights the very issue we raised in Chapter 4: trajectory matching can be a misleading proxy for true physical understanding. A model can be optimized to produce trajectories that are geometrically closer to a single ground truth (i.e., lower L2 error) without necessarily capturing the underlying physical laws as effectively. Since our goal is to instill a generalizable understanding of physics, we prioritize the findings from our explicit Morpheus evaluation.

While these results suggest a clear, proxy-dependent preference for the loss function, we refrain from over-generalizing. The optimization landscape for these objectives is highly complex, and it is plausible that one loss type is simply easier to optimize for a given teacher signal. Nonetheless, our results clearly show that both alignment strategies are highly effective, with the optimal choice depending on the specific physics proxy being distilled.

Challenges in Multi-Proxy Alignment A core strength of PIRA is its flexibility in using diverse physics proxies. This motivated us to explore whether combining multiple proxy physics-based transformed signals (e.g, depth, masks, and optical flow) into a single training objective could yield a more holistically physics-aware VDM. In a preliminary experiment, we formulated the total distillation loss as an unweighted sum of the individual $\mathcal{L}_{mdms}$ (Equation 3.6) components for each proxy. However, as illustrated by the training curves in Figure A.2, this approach led to pathological training dynamics. A similar "tug-of-war" between loss components arose when we employed the $\mathcal{L}_{mcos}$ objective (Equation 3.3) instead; as this did not reveal any new patterns, we find it redundant to present those graphs here for brevity.

Our analysis of the individual loss components reveals a clear conflict between optimization goals. The loss for optical flow, a proxy for 2D dynamics, minimizes rapidly from over 0.2 to approximately 0.02 (Figure A.2b). In stark contrast, the loss for depth, which encodes static 3D geometry, increases dramatically from ≈ 0.85 to a peak above 1.05 (Figure A.2d). The mask loss exhibits a compromise, dropping sharply before rebounding and stagnating (Figure A.2c). This suggests that the steep, easily attainable gradients from the dynamically-oriented tasks guide the optimizer into a region of the loss landscape into a local minimum, favorable for motion but fundamentally incompatible with aligning to the geometric structure of the depth space. While the total distillation loss converges to a seemingly stable value around 1.1 (Figure A.2a), this conceals an underlying problem where the model masters one task by compromising another one. This finding indicates that while PIRA is a promising framework, its extension to multi-proxy scenarios is a non-trivial multi-task learning problem requiring careful consideration of its training dynamics. Such an endeavor would likely necessitate techniques such as manual loss weighting or curriculum learning, as discussed in Chapter 8.

Combining Alignment Objectives Given that both alignment strategies(See Section 3.3) demonstrated strong, proxy-dependent performance, we investigated whether a multi-loss objective combining them could yield further improvements. To navigate the complex hyperparameter manifold, our multi-loss configuration for each proxy was constructed by summing the individually best-performing point-to-point (Equation 3.3) and relational (Equation 3.6) loss setups. As detailed in Table A.4, this *multi-loss* approach consistently underperformed compared to the best single-loss variant for each proxy. The performance degradation was particularly pronounced for the Depth-based proxy, where the multi-loss configuration yielded a Combined Score of **0.850** in the ID split, a considerable drop of over 7 percentage points from the best single-loss score of 0.921. This substantial gap strongly suggests a *tug-of-war* dynamic similar to the one observed in our multi-proxy alignment; the distinct optimization pressures of the point-to-point and relational objectives appear to create conflicting gradients, resulting in an inefficient optimization landscape. This finding reinforces our conclusion that a single, well-chosen loss function is more effective and flexible than combining objectives with competing gradients.

The Role of the Projection MLP in Relational Alignment Our framework allows for the optional inclusion of a trainable projection MLP before the Distance Matrix Similarity loss (Section 3.3.2 - Eq. 3.6), a component that is necessary for carrying out the representation alignment with the point-to-point loss(Section 3.3.1), to align feature dimensions between teacher (transformed physics-grounded proxy signals) and student (noisy hidden states of VDM) features. A consistent and significant finding across

all our experiments is that the best-performing relational alignment configurations are achieved *without* a trainable MLP. This pattern holds true across both implicit and explicit evaluation metrics. For instance, in the explicit *Morpheus* evaluation (Table A.5), the top-performing Optical Flow method (MDMS No MLP, $m = 0.1$) achieves the best OOD Combined score of *0.928*. The same trend is observed for the other proxies, where the *No MLP* variants consistently outperform their *MLP* counterparts. This strongly validates our core architectural choice: re-purposing the VDM's native 3D VAE as the physics encoder creates teacher representations whose feature space is already inherently compatible with the VDM's internal states. The results suggest that an additional learned projection is not only redundant for achieving strong relational alignment but can even degrade performance by complicating the optimization landscape. Our finding simplifies the overall PIRA framework and indicates the redundancy of training an extra network, in contrast to prior works like VideoREPA Zhang et al. (2025b) which assume it as a necessity.

Sensitivity to the Margin Hyperparameter Our framework incorporates a margin hyperparameter, m, in both representation-alignment objective functions (Section 3.3) to prevent the model from being over-penalized for already well-aligned features. Our ablations (Table A.5) reveal that the optimal margin interacts strongly with both the loss function and the teacher signal. For the *point-to-point objective* (Section 3.3.1), the optimal margin varies dramatically: the top-performing Mask-based model benefits from a large margin of $m = 0.5$, whereas the best Depth-based implicit performance requires a strict zero-margin, $m = 0.0$ (Table A.1). This high variance suggests that the ideal degree of tolerance for direct feature alignment is highly task-dependent.

In contrast, a clear and consistent pattern emerges for the *relational* objective(Section 3.3.2): a small, non-zero margin is almost always superior. This is strikingly evident in the TORA-based approach, where increasing the margin from $m = 0.0$ to $m = 0.1$ improves the OOD Combined Score by nearly three percentage points, from **0.880** to **0.900**. This pattern suggests that while relational alignment benefits from a small amount of "slack" to focus on the most significant structural differences, the point-to-point objective's behavior is more sensitive. The overall strong performance achieved across this wide range of configurations validates the general robustness of PIRA.

Chapter 7

Conclusion

In this work, we confronted the *physics gap* in state-of-the-art Video Generative Models, where impressive visual fidelity in generation often contradicts adherence to fundamental physical dynamics. Our proposed framework, **PIRA**, closes this gap by grounding Video Diffusion Models in the principles of the physical world. The virtue of our approach lies in its simple yet effective architectural design: we redefine the VDM's own native 3D VAE to be used as an inherently compatible encoder for first-principles physics proxies. This elegant solution bypasses the latent space mismatch that complicates other knowledge distillation methods, enabling a stable and generalizable fine-tuning process.

Through extensive experimentation on the simple yet challenging task of modeling free-fall dynamics, we have validated PIRA's empirical superiority. Our results show that PIRA significantly outperforms both standard fine-tuning and prior alignment-based methods across a comprehensive suite of metrics. The most profound insight, however, is not just *that* PIRA performs better, but *why*. While implicit state variable matching confirms PIRA's ability to produce more accurate outcomes, it is our explicit, physics-informed scores that unveil the deeper mechanism. This analysis reveals that PIRA is uniquely effective at teaching the model to respect the underlying governing physical dynamics, rather than merely conserving select physical quantities. This distinction is critical, as it suggests that for generative models to serve as accurate world simulators, they must be grounded in the causal, dynamical principles of the world.

Implications The findings presented here have broader implications for the future of physics-aware generative modeling. We have demonstrated that a *proxy-based physics representation target* grounded in interpretable, first-principles signals—offers a more direct and practical approach to instilling genuine physical plausibility. The success and simplicity of our methodology suggest a promising direction for creating more reliable and trustworthy generative models. By providing a stable method for bridging the gap between visual realism and physical accuracy, this work represents a substantial step toward realizing the vision of VDMs as faithful world models and transforming them into reasoners of the physical world in which we live.

Chapter 8

Limitations & Future Work

While our results demonstrate a significant step forward, it is essential to acknowledge the limitations of this work and outline the exciting avenues for future research that these limitations reveal.

8.1 Limitations

Scope of Physical Phenomena Our experimental validation was intentionally focused on a single, fundamental physical phenomenon: free-fall dynamics under normal Earth's gravity. This choice provided a controlled and rigorous test case to demonstrate the efficacy of our method clearly. Yet, while the PIRA framework is designed to be general, its performance on more complex physical systems has not been empirically validated. Testing on a set of experiments that do not follow linear dynamics, multi-object interactions, and other complex physical systems would most likely cause our method to break, and would require re-thinking and redesigning a framework for grounding Video Diffusion Models with that in mind.

Reliance on Physics Proxies PIRA operates by distilling knowledge from observable signals that can be interpreted as consequences of physical laws, rather than by directly incorporating the governing equations into the model's architecture (e.g., as in Physics-Informed Neural Networks). The quality of the learned physical understanding, therefore, depends on the fidelity and completeness of these proxies. For instance, optical flow estimators can be noisy or fail in texture-less regions, and segmentation masks do not capture crucial intrinsic properties like mass, friction, or the coefficient of restitution. This reliance on observable proxies means that the distilled knowledge is at best an approximation of the true physical state.

Synthetic-to-Real Generalization The development and validation of our framework were conducted on a synthetic dataset. As argued in Chapter 5, this was a principled choice that enabled controlled experimentation and access to perfect, noise-free ground-truth annotations, which is essential for cleanly isolating the effects of our method. However, the successful transfer to the PIRA-enhanced model to effectively generate or reason about the dynamics in noisy, complex, and unconstrained real-world videos presents an additional generalization challenge, which is one of the first things we aim to test next.

Challenges in Multi-Proxy Alignment A core strength of PIRA is its flexibility in using diverse signals that act as physics proxies. However, our preliminary experiments in combining multiple physics-based transformations originating from our proxy signals (e.g masks, depth, and optical flow) into a single training objective revealed a significant challenge inherent to multi-task learning. Specifically the unweighted sum of the individual alignment losses (one-per proxy signals) resulted in a "tug-of-war" between conflicting optimization goals. We attribute this to the fact that gradients from easier-to-learn, dynamically-oriented tasks (e.g., optical flow) dominated the learning process, actively hindering or even reversing progress on the more difficult geometric task of aligning to depth features. This indicates that while PIRA is a promising framework, its extension to multi-proxy scenarios is not a trivial plug-and-play operation and requires careful consideration of the training dynamics. Successfully balancing these competing objectives would likely necessitate further techniques such as *manual loss weighting*, to amplify the signal from harder tasks, or a *curriculum learning* strategy, where the model is trained sequentially on proxies of increasing difficulty.

8.2 FUTURE WORK

Given the limitations of this work, we naturally define a clear roadmap for future research:

From Proxy Physics-Teachers to Explicit Physics Encoders Another promising direction is to explore the design of an explicit physics-based teacher network. Rather than relying on proxies like optical flow, a dedicated physics-encoder model could be trained to estimate explicit kinematic state variables—such as position, velocity, and rotation—directly from video frames. A concurrent work[1], NewtonGen (Yuan et al., 2025), provides a compelling minimal viable working design for this approach. A *Neural Newtonian Dynamics* (NND) model is proposed, which consists of a mix of General Physics-Informed Neural ODEs (Chen et al., 2018), to predict future physical states, which are then used to condition a separate video generator. This validates the potential of using an explicit, learnable physics engine as a teacher. Future work would investigate using our representation alignment framework to *distill* the knowledge from such a sophisticated NND model into a VDM in an end-to-end fashion. This could potentially simplify the two-stage pipeline of NewtonGen while leveraging PIRA's stable alignment mechanism. Furthermore, hybrid strategies could be explored, like fusing the rich semantic context from large VFMs with the explicit dynamical predictions from a physics-centric model like an NND, potentially offering the best of both worlds.

Bridging the Sim-to-Real Gap A crucial next step is to bridge the gap between synthetic training and real-world application. In the first step, we plan to test the zero-shot capabilities of PIRA on a curated dataset of real-world physical experiments. That will include testing our method on the falling items experiment of our Morpheus benchmark (Zhang et al., 2025a) and on the real-world split of PisaBench (Li et al., 2025), which includes 361 videos of falling items. Such experiments would rigorously test the robustness and generalization of the instilled physical knowledge, assessing how well the learned principles transfer from a clean, simulated environment to actual real-world videos.

Expanding the Physical Domain A primary avenue for future work is to apply the PIRA framework to a wider range of physical phenomena. This would involve creating new synthetic datasets for other fundamental Newtonian dynamics or using the experiments recorded in Morpheus (Zhang et al., 2025a) and their augmentations obtained with Cosmos-Transfer (Alhaija et al., 2025) video-to-video stylization. This includes experiments such as projectile motion, collisions, and pendulum mechanics.

[1] Released on 25^{th} of September.

Bibliography

Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025.

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13–23, 2025.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Ian J 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.

Google. Veo 3 Model Card. https://aistudio.google.com/models/veo-3, 2024. Accessed: 2024-10-27.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, et al. " phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Sungwon Hwang, Hyojin Jang, Kinam Kim, Minho Park, and Jaegul Choo. Cross-frame representation alignment for fine-tuning video diffusion models. *arXiv preprint arXiv:2506.09229*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, Ajmal Mian, Mubarak Shah, and Chang Xu. Generative physical ai in vision: A survey. *arXiv preprint arXiv:2501.10928*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4040–4048, 2016.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4991, 2023.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017a.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part ii): Data-driven discovery of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10566*, 2017b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Kun Xiang, Terry Jingchen Zhang, Yinya Huang, Jixi He, Zirong Liu, Yueling Tang, Ruizhe Zhou, Lijing Luo, Youpeng Wen, Xiuwei Chen, et al. Aligning perception, reasoning, modeling and interaction: A survey on physical ai. *arXiv preprint arXiv:2510.04978*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025.

Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025b.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025c.

Appendix A

Appendix

A.1 Details of Motion-Specialized Encoder Alignment

As a point of comparison for our main PIRA framework, we implemented an alternative alignment strategy based on a motion-specialized encoder, inspired by the architecture of Tora (Zhang et al., 2025c). ora was originally introduced to provide explicit motion control by *fusing* user-specified trajectories directly into the VDM during inference. To achieve this, it jointly trains a Motion-Guidance Fuser (MGF) with a **Trajectory Extractor (TE)** network. The TE's sole purpose is to convert an external motion signal, namely an optical flow field, into a set of dense, patch-based latent representations (*motion patches*) that are directly compatible with the VDM's internal features.

The TE encoder, is illustrated in Figure A.1, is a sophisticated motion encoder. It processes optical flow field predictions from estimator models like (Teed & Deng, 2020) representing pixel-wise motion vectors between consecutive video frames [1], $g_{map} \in \mathbb{R}^{F \times H \times W \times 2}$. This vector is first compressed by a specialized 3D-VAE, which itself was finetuned from MAGVIT-v2 (Yu et al., 2023) on a combination of datasets with optical-flow annotations (Cabon et al., 2020; Mehl et al., 2023; Mayer et al., 2016). Due to the scarcity of a generic optical-flow compression network, the MAGVIT-V2 was re-purposed and finetuned to preserve the motion bias of the input signal. The resulting compressed latent $g_{vae} \in \mathbb{R}^{f \times h \times h \times d}$ is then processed by a series of convolution blocks [2] to produce hierarchical spacetime motion patches $y \in \mathbb{R}^{l \times s \times d'}$ [3]. By re-purposing this pre-trained TE, we can create dense, VDM-compatible alignment targets $\mathbf{y}^{\star}_{\mathbf{physics}}$ directly from optical flow. We then apply the representation alignment loss, inspired by VideoREPA (Zhang et al., 2025b), to softly align the VDM's intermediate features $\mathbf{h_t}$ with the derived motion-patches. This TORA-based approach successfully combines a first-principles physics proxy (optical flow) with a stable, alignment mechanism.

[1] such vectors contain the 2D displacement vectors (dx, dy) for each pixel location.

[2] Subsequently, as many convolution blocks as DiT layers in VDM, the output of the $i-th$ convolution produces a motion-patch to be fused with the corresponding VDM hidden state, which are the input to the MGF component.

[3] Any dimensionality mismatch between VFM $\mathbf{h_{student}}$ latent noisy states and encoded motion-patches $\mathbf{y_{student}}$ is resolved when applying the temporally consistent representation-alignment losses.

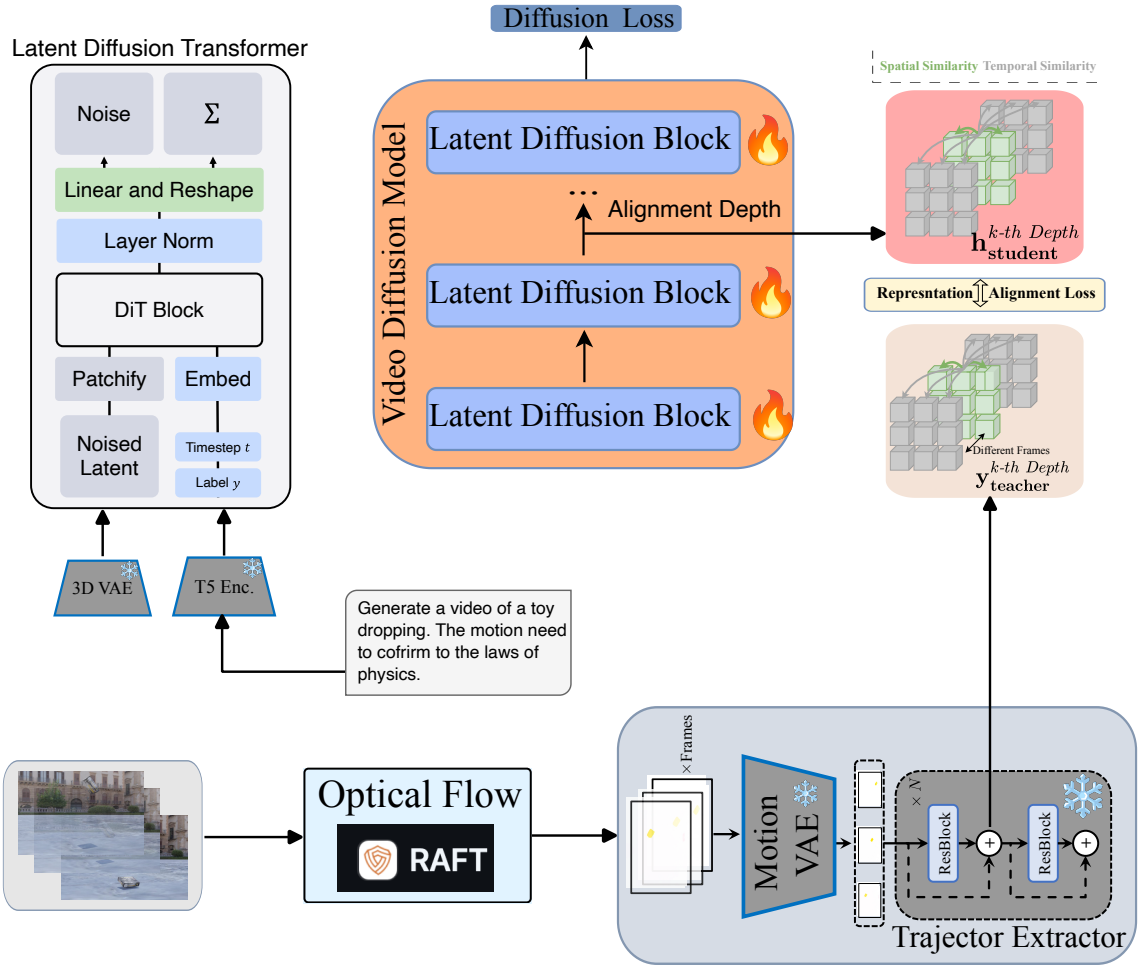


Figure A.1: The TORA-based representation alignment pipeline. An optical flow map is processed by the Trajectory Extractor (TE) to produce VDM-compatible motion patches. These patches serve as the teacher representation target to carry out the alignment, which regularizes an intermediate layer of the main VDM.

A.2 Comprehensive Results

Method		ID		OOD	
		L2 (↓)	CD (↓)	L2 (↓)	CD (↓)
Oracle	SAM-2 GT	0.001	0.000	0.000	0.000
Baselines	CogVideoX-5B-I2V	0.048	0.123	0.052	0.130
	SFT CogVideoX-5B-I2V	0.028	0.070	0.021	0.046
TORA based	PIRA ($\mathcal{L}_{mdms}$ m=0.05)	0.024	0.058	0.021	0.044
	PIRA ($\mathcal{L}_{mdms}$ m=0.1)	0.020	0.048	0.019	0.036
	PIRA ($\mathcal{L}_{mdms}$ m=0.0)	0.020	0.044	0.020	0.040
Depth Methods	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.019	0.042	0.018	0.035
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.018	0.040	0.018	0.035
	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.0)	0.023	0.055	0.020	0.044
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.019	0.043	0.016	0.030
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.015	0.031	0.013	0.022
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.021	0.048	0.019	0.038
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.019	0.041	0.017	0.032
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.022	0.052	0.019	0.040
Mask Methods	Multi-Loss MLP (MCCOS=0.5, MDMS=0.1)	0.017	0.033	0.016	0.029
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.016	0.031	0.014	0.025
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.014	0.027	0.014	0.023
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.017	0.033	0.015	0.025
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.016	0.033	0.014	0.024
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.017	0.037	0.016	0.027
	MCCOS ($\mathcal{L}_{mccos}$ m=0.5)	0.012	0.020	0.013	0.021
OpticalFlow Methods	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.1)	0.019	0.041	0.016	0.031
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.023	0.057	0.021	0.045
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.019	0.043	0.015	0.027
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.013	0.022	0.014	0.024
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.015	0.030	0.014	0.024
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.014	0.027	0.016	0.031

Table A.1: PisaBench-style evaluation for trajectory and shape fidelity across all ablation studies. We highlight the best and second-best values for each metric. The Oracle row represents the ground truth and is excluded from the comparison. (↓) Lower is better.

Method		ID			OOD		
		L2(↓)	**MSE (var)(↓)**	**MSE (spat)(↓)**	**L2(↓)**	**MSE (var)(↓)**	**MSE (spat)(↓)**
Oracle	SAM-2 GT	0.001	0.000	0.000	0.000	0.000	0.000
Baselines	CogVideoX-5B-I2V	0.126	0.637	0.451	0.096	0.697	0.455
	SFT CogVideoX-5B-I2V	0.033	0.304	0.244	0.027	0.214	0.197
TORA based	PIRA ($\mathcal{L}_{mdms}$ m=0.05)	0.033	0.247	0.206	0.029	0.218	0.190
	PIRA ($\mathcal{L}_{mdms}$ m=0.1)	0.025	0.204	0.166	0.019	0.186	0.160
	PIRA ($\mathcal{L}_{mdms}$ m=0.0)	0.020	0.191	0.156	0.020	0.206	0.176
Depth Methods	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.019	0.165	0.156	0.018	0.178	0.154
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.018	0.151	0.147	0.018	0.158	0.153
	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.0)	0.039	0.235	0.198	0.054	0.231	0.190
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.019	0.162	0.158	0.018	0.143	0.131
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.016	0.114	0.110	0.013	0.097	0.094
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.037	0.203	0.181	0.033	0.183	0.162
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.022	0.154	0.156	0.017	0.135	0.136
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.026	0.203	0.180	0.020	0.197	0.161
Mask Methods	Multi-Loss MLP (MCCOS=0.5, MDMS=0.1)	0.018	0.133	0.141	0.016	0.136	0.130
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.016	0.121	0.124	0.014	0.112	0.107
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.015	0.108	0.102	0.014	0.117	0.105
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.017	0.130	0.137	0.015	0.115	0.113
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.016	0.125	0.134	0.014	0.107	0.106
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.020	0.144	0.140	0.016	0.122	0.123
	MCCOS ($\mathcal{L}_{mccos}$ m=0.5)	0.012	0.072	0.078	0.013	0.092	0.088
OpticalFlow Methods	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.1)	0.019	0.161	0.151	0.016	0.140	0.130
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.023	0.233	0.197	0.022	0.222	0.182
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.020	0.163	0.155	0.015	0.123	0.113
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.013	0.083	0.083	0.014	0.101	0.103
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.022	0.121	0.121	0.016	0.116	0.108
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.023	0.111	0.111	0.016	0.130	0.128

Table A.2: Ablation study on occlusion-robust metrics. We highlight the best and second-best values for each metric. The Oracle and Baselines are excluded from the comparison. (↓) Lower is better.

Method		ID			OOD		
		ST-IoU (↑)	**S-IoU (↑)**	**W-S-IoU (↑)**	**ST-IoU (↑)**	**S-IoU (↑)**	**W-S-IoU (↑)**
Oracle	SAM-2 GT	0.911	0.921	0.911	0.931	0.947	0.931
Baselines	CogVideoX-5B-I2V	0.235	0.403	0.314	0.235	0.398	0.308
	SFT CogVideoX-5B-I2V	0.391	0.594	0.546	0.432	0.651	0.594
TORA based	PIRA ($\mathcal{L}_{mdms}$ m=0.05)	0.412	0.608	0.557	0.445	0.658	0.591
	PIRA ($\mathcal{L}_{mdms}$ m=0.1)	0.460	0.616	0.577	0.473	0.645	0.604
	PIRA ($\mathcal{L}_{mdms}$ m=0.0)	0.466	0.595	0.570	0.455	0.629	0.587
Depth Methods	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.458	0.624	0.585	0.456	0.650	0.599
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.471	0.628	0.591	0.463	0.666	0.615
	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.0)	0.437	0.607	0.564	0.444	0.642	0.591
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.464	0.626	0.587	0.505	0.662	0.612
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.493	0.629	0.596	0.524	0.688	0.640
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.423	0.612	0.574	0.464	0.646	0.600
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.470	0.640	0.599	0.482	0.675	0.624
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.434	0.596	0.559	0.475	0.650	0.600
Mask Methods	Multi-Loss MLP (MCCOS=0.5, MDMS=0.1)	0.471	0.623	0.591	0.488	0.678	0.625
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.482	0.640	0.606	0.521	0.678	0.632
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.514	0.639	0.608	0.505	0.668	0.626
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.469	0.632	0.596	0.507	0.683	0.639
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.474	0.638	0.603	0.514	0.685	0.639
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.469	0.623	0.588	0.497	0.677	0.630
	MCCOS ($\mathcal{L}_{mccos}$ m=0.5)	0.525	0.655	0.625	0.517	0.684	0.646
OpticalFlow Methods	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.1)	0.459	0.612	0.575	0.489	0.671	0.618
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.432	0.600	0.558	0.457	0.642	0.593
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.445	0.610	0.571	0.500	0.676	0.628
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.523	0.643	0.612	0.494	0.674	0.633
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.480	0.625	0.595	0.502	0.666	0.626
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.491	0.636	0.603	0.489	0.681	0.630

Table A.3: Ablation study on Physics-IQ metrics. We highlight the best and second-best values for each metric. The Oracle and Baselines are excluded from the comparison. (↑) Higher is better.

Method		ID			OOD		
		Dyn. (↑)	**Phys. (↑)**	**Comb. (↑)**	**Dyn. (↑)**	**Phys. (↑)**	**Comb. (↑)**
Oracle	SAM-2 GT	0.998	0.944	0.971	0.959	0.953	0.956
Baselines	CogVideoX-5B-I2V	0.392	0.638	0.515	0.400	0.743	0.572
	SFT CogVideoX-5B-I2V	0.672	0.782	0.727	0.636	0.822	0.729
TORA based	Specialized Optical-flow Bias	0.897	0.901	0.899	0.889	0.911	0.900
PIRA	Depth-based Physics (Best)	0.930	0.912	0.921	0.928	0.912	0.920
	Depth-based Physics (Multi-Loss)	0.880	0.819	0.850	0.832	0.821	0.827
	Mask-based Physics (Best)	0.995	0.939	0.967	0.902	0.939	0.921
	Mask-based Physics (Multi-Loss)	0.961	0.897	0.929	0.897	0.921	0.909
	Optical Flow-based Physics (Best)	0.979	0.926	0.953	0.917	0.938	0.928
	Optical Flow-based Physics (Multi-Loss)	0.960	0.894	0.927	0.869	0.922	0.896

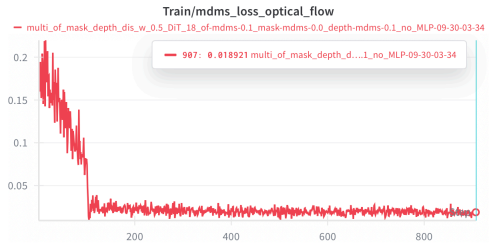
Table A.4: Morpheus physics-based evaluation comparing best and multi-loss models. We highlight the best and second-best values for each metric across all competing methods (TORA and PIRA). Column headers are abbreviated: Dyn. (Dynamical), Phys. (Physical), and Comb. (Combined). Oracle and Baselines are excluded from the comparison. (↑) Higher is better.

Method		ID			OOD		
		Dyn. (↑)	**Phys. (↑)**	**Comb. (↑)**	**Dyn. (↑)**	**Phys. (↑)**	**Comb. (↑)**
Oracle	SAM-2 GT	0.998	0.944	0.971	0.959	0.953	0.956
Baselines	CogVideoX-5B-I2V	0.392	0.638	0.515	0.400	0.743	0.572
	SFT CogVideoX-5B-I2V	0.672	0.782	0.727	0.636	0.822	0.729
TORA based	PIRA ($\mathcal{L}_{mdms}$ m=0.05)	0.868	0.862	0.865	0.836	0.897	0.867
	PIRA ($\mathcal{L}_{mdms}$ m=0.1)	0.875	0.889	0.882	0.889	0.911	0.900
	PIRA ($\mathcal{L}_{mdms}$ m=0.0)	0.897	0.901	0.899	0.856	0.904	0.880
Depth Methods	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.944	0.890	0.917	0.869	0.924	0.897
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.913	0.833	0.873	0.865	0.873	0.869
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.930	0.912	0.921	0.928	0.912	0.920
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.929	0.898	0.914	0.883	0.917	0.900
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.953	0.896	0.925	0.866	0.927	0.897
	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.0)	0.880	0.819	0.850	0.832	0.821	0.827
	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.952	0.884	0.918	0.899	0.933	0.916
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.930	0.900	0.915	0.883	0.937	0.910
Mask Methods	MCCOS ($\mathcal{L}_{mccos}$ m=0.5)	0.995	0.939	0.967	0.902	0.939	0.921
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.977	0.907	0.942	0.900	0.939	0.920
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.977	0.916	0.947	0.900	0.916	0.908
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.962	0.866	0.914	0.900	0.934	0.917
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.968	0.923	0.946	0.900	0.922	0.911
	Multi-Loss MLP (MCCOS=0.5, MDMS=0.1)	0.961	0.897	0.929	0.897	0.921	0.909
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.977	0.919	0.948	0.900	0.930	0.915
OpticalFlow Methods	MCCOS ($\mathcal{L}_{mccos}$ m=0.0)	0.960	0.872	0.916	0.885	0.929	0.907
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.0)	0.913	0.896	0.905	0.822	0.929	0.876
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.05)	0.953	0.892	0.923	0.902	0.923	0.913
	MDMS MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.962	0.881	0.922	0.870	0.929	0.900
	Multi-Loss No MLP (MCCOS=0.0, MDMS=0.1)	0.960	0.894	0.927	0.869	0.922	0.896
	MDMS No MLP ($\mathcal{L}_{mdms}$ m=0.1)	0.979	0.926	0.953	0.917	0.938	0.928

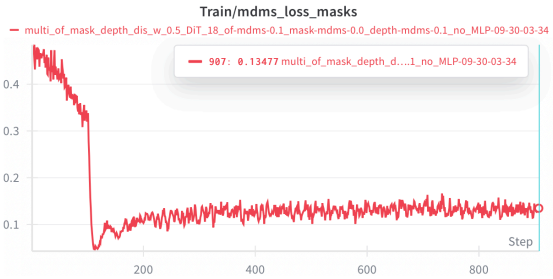
Table A.5: Ablation study on Morpheus physics-based metrics. We highlight the best and second-best values for each metric across all competing methods. Column headers are abbreviated: Dyn. (Dynamical), Phys. (Physical), and Comb. (Combined). Oracle and Baselines are excluded from the comparison. (↑) Higher is better.

A.3 Combining Different Physics-Proxy bias for Representation Alignment

(a) Total Distillation Loss



(b) $\mathcal{L}_{mdms}$ for Optical Flow



(c) $\mathcal{L}_{mdms}$ for Masks

(d) $\mathcal{L}_{mdms}$ for Depth

Figure A.2: Training loss curves for a multi-proxy PIRA model with an unweighted sum of losses, revealing a "tug-of-war" between conflicting objectives. While the total loss (a) appears to converge, the individual components show that the easily optimized optical flow loss (b) dominates, while the depth loss (d) is actively driven higher. This indicates that the conflicting gradients create a complex loss landscape, preventing the model from finding a minimum that jointly satisfies all physics-proxy alignment tasks.

Weakness 1.1: Limited Scope of Physical Phenomena "The experimental validation was intentionally restricted to a single..."

Answer 1.1: I acknowledge that extending the work to broader Newtonian dynamics would provide valuable insights regarding the method's ability to generalize. As correctly pointed out, this was a deliberate design choice; however, I posit that the current scope of the proposed method is not merely a starting point, but a necessary step to address the gap in instilling physics knowledge into pre-trained video models.

Fine-tuning multiple model variants (including depth, masks, and optical flow, along with two representation alignment losses and their combinations) covering a broad set of hyper-parameters to assess sensitivity and extract the best performance—along with baselines—on a large-scale synthetic dataset of 10k samples placed a significant strain on the available compute budget. For instance, a single fine-tuning run takes approximately 20 hours on 4 H100 GPUs. This does not account for the resources needed for validation on the extensive runs and datasets we have (Implicit Physical Plausibility Quantification via State Variable Matching, Physics-IQ, PISA, MSE-devised Metrics, and the Explicit Physical Plausibility Quantification benchmark). We have two validation splits, as described in the paper (In-Distribution and Out-Of-Distribution), to carefully study the generalization of our method.

Future plans will include extending the benchmark to other types of Newtonian dynamics, such as those covered in the Morpheus Benchmark.

Weakness 1.2: Limited Scope of Only Synthetic Data Training "The successful transfer of the PIRA-enhanced model to generate or reason about dynamics in noisy, complex, and unconstrained real-world videos remains an untested generalization challenge."

Answer 1.2: To test the generality of PIRA and bridge the sim-to-real gap, I have devised a strategy using the PISA-bench dataset, which includes a split of 360 video samples of real-world falling item experiments. Here are the actions and steps taken so far to incorporate this into the evaluation analysis of our proposed methodology:

Incompatibility of resolution and duration: The real data has an aspect ratio of 1:1 and different FPS settings, in contrast to our synthetic data, which explicitly matched the CogVideo-5B-I2V backbone settings. To resolve this challenge while following the PISA work as closely as possible, we took the following actions: i) For all baselines and our method that do not support generating videos with a 1:1 aspect ratio, we pad the initial frames with black borders to reach the resolution supported by these models, and finally remove the borders from the generated videos. Thus, the resulting video has a 1:1 ratio. The initial padding requires careful consideration as the original real-world video resolution is 1080x1080. ii) Frame Rate discrepancy between generated and ground-truth videos: To perform FPS alignment, we map each frame index of the generated videos to the ground truth using in the ground truth. Resource usage: Generating 360 videos per comparison type requires significant computational resources, particularly when accounting for cluster resource management. Current Status: We have acquired results for our best-performing PIRA variants, which are encouraging, and we plan to update the results on the thesis-dedicated website: https://physics-informed-repa.github.io/PIRA-website/. Due to resource constraints—as generating videos with such models takes time—the comparison with baselines will be provided later. This is especially relevant considering that we provide both pixel-level evaluation (trajectory-level metrics) and the quantification of physical reasoning in the MORPHEUS benchmark (dynamical scores), which requires training a PINN per generated video.

Weakness 2: Challenges with Multi-Proxy Alignment "Preliminary experiments combining multiple physics-based proxy signals... revealed a significant challenge characteristic of multi-task learning... Potentially something like Agglomerative distillation should be applied..."

Answer 2: As correctly pointed out in our ablation studies, combining different types of signals requires a smarter design. Upon the supervisor's request, in a future version, we plan to follow Agglomerative Distillation as described in previous works (RadioV2.5: https://arxiv.org/pdf/2412.07679). This approach has been tested on state-of-the-art self-supervised encoders (like SAM) but on the latent space of video diffusion models, posing a smart design for carrying out representation alignment with multiple targets.

Another choice to unlock better performance would be to attempt aligning the proxy-physics bias in different DiT hidden layers. Rather than conducting a massive ablation study, we could design a heuristic to reveal the compatibility between the noisy hidden states of the VDM and the encoded proxy-physics signals. It would be interesting to see if this resolves the competing gradient problems from which PIRA currently suffers.

Weakness 3: "Comparison with DINO-v2 based video alignment would further improve proper presentation of the work"