

RESEARCH PROPOSAL

Synopsis

For

Zero-Shot Sim-to-Real Physics Inference from Video Foundation Models

BIPIN JOSEPH
24MCA107

MCA Artificial Intelligence and Machine Learning with Minor in Data Science
The Yenepoya Institute of Arts, Science, Commerce & Management
Mangalore

27/10/2025

Under the guidance of
Dr Parameshwar R Hegde
Assistant Professor, Department of Computer Science
The Yenepoya Institute of Arts, Science, Commerce & Management
Mangalore

Submitted to



**THE YENEPOYA INSTITUTE OF ARTS, SCIENCE, COMMERCE &
MANAGEMENT (YIASCM)
YENEPOYA (DEEMED TO BE UNIVERSITY)**

MANGALORE, KARNATAKA

I. Title of the Project

“Zero-Shot Sim-to-Real Physics Inference from Video Foundation Models”

II. Introduction

The accurate simulation of physical dynamics is a cornerstone of realistic digital content creation (e.g., gaming, VFX) and robust autonomous systems (e.g., robotics). A significant bottleneck in these domains is the **parameterization** of physical models—assigning quantitative values for properties like **elasticity** or **dynamic friction**, which is often performed manually through iterative refinement. Recent advances in **Video Foundation Models (VFM**s)—large-scale neural networks pre-trained on diverse video datasets using objectives like generative modeling or self-supervised learning—have demonstrated remarkable capabilities in semantic video understanding. These models implicitly learn rich, high-dimensional **representations** of motion and dynamics. This project investigates the potential of leveraging these learned representations for **quantitative physical inference**: extracting explicit, continuous-valued physical parameters directly from video input, thereby automating a critical aspect of physically-based modeling.

III. Current Trends in the Topic

Research into visual physics inference spans several paradigms. **Differentiable physics engines** enable gradient-based **system identification** but are computationally intensive and require strong priors. **Heuristic-based methods**, often termed "oracle" approaches, use classical computer vision to track explicit visual cues (e.g., trajectories, deformation) but lack robustness and generalizability. Unsupervised learning via **generative models** (e.g., next-frame prediction) captures **latent dynamics** but yields implicit representations unsuitable for direct quantitative querying. The current trend involves leveraging large-scale **foundation models**. While benchmarks like Physion++ assess **qualitative reasoning**, the work by Zhan et al. (2025) pioneers probing VFMs (both generative like DynamiCrafter and self-supervised like V-JEPA-2) for

quantitative values. However, this work critically highlights the **simulation-to-reality (Sim-to-Real) gap** as a major impediment to real-world applicability, necessitating domain adaptation techniques like fine-tuning.

IV. Aim of the Research

The primary aim of this research is to **systematically investigate and quantify the implicit physical knowledge encoded within the latent representations of pre-trained Video Foundation Models**. Specifically, it seeks to develop and rigorously evaluate a methodology for extracting explicit, quantitative estimates of dynamic physical properties (elasticity, friction) from raw video by probing frozen VFM feature hierarchies. A central objective is to **address the Sim-to-Real generalization gap** by exploring the efficacy of **domain randomization** during synthetic data generation. The goal is to achieve robust **zero-shot inference** on real-world videos, thereby obviating the need for fine-tuning on target domain data, which was a limitation in prior work .

V. Problem statement

The core problem is the inefficiency and lack of empirical grounding in current workflows for parameterizing physically-based simulations. Existing automated physics inference methods are either domain-specific, computationally prohibitive, or yield non-interpretable results. While VFMs offer a promising avenue due to their rich learned representations of dynamics, extracting **quantitative** physical parameters remains an "underexplored" challenge. Furthermore, models trained purely on synthetic data "struggle to generalize" to real-world videos, requiring domain adaptation. This leads to the specific research question: **Can domain randomization applied during synthetic video generation enable a VFM-based probing architecture (utilizing a frozen backbone and a lightweight trainable head) to perform accurate quantitative physical inference on real-world videos in a zero-shot manner, significantly mitigating the Sim-to-Real gap observed in baseline approaches?**

VI. Research Methodology

The research will employ a structured, empirical methodology:

1. Synthetic Dataset Generation with Domain Randomization: A labelled dataset (PhysVid-style) will be generated using the Blender physics engine and its Python

scripting API (bpy). Videos depicting canonical physical scenarios will be created. Crucially, domain randomization will be applied by programmatically varying nuisance parameters per video, including camera viewpoint (pose and potentially intrinsics), object appearance (randomized PBR textures, albedo), lighting conditions (HDRI maps, light source properties), and initial object states, following principles outlined in [arXiv:2510.02311's](#) appendix . This forces the model to learn representations invariant to visual style.

2. Model Architecture: The inference model employs a frozen VFM backbone (e.g., V-JEPA-2, a ViT-based architecture) accessed via Hugging Face transformers. A lightweight, trainable probing module implemented in PyTorch will perform the inference task. This module comprises:
 - A learnable query vector q : This vector parameterizes the specific physical property being queried.
 - A cross-attention mechanism: This compute attention scores between q and the VFM's spatiotemporal feature tokens (r_i) from multiple layers, allowing q to selectively pool relevant features: $P = \Sigma \text{softmax}(q \cdot r_i) \cdot r_i$. *This isolates task-specific information from the general-purpose VFM features.*
 - An MLP Regression Head: A multi-layer perceptron maps the aggregated feature vector P to the final scalar prediction.
3. Training Protocol: Only the probing module parameters will be optimized using supervised learning on the synthetic dataset via the Adam optimizer. Following Zhan et al. (2025) , the loss function will be L1 loss for elasticity/friction and Log L1 loss for viscosity (to handle its large dynamic range).
4. Evaluation Protocol: Model performance will be quantitatively assessed using the Pearson Correlation Coefficient on distinct test splits: test-1 (in-distribution synthetic), test-2 (out-of-distribution synthetic), and test-3 (real-world videos) . Pearson correlation effectively measures the ordinal correctness and linear relationship, crucial for evaluating generalization across domains where absolute scales might shift. Success criteria include high correlation on synthetic data and significantly improved zero-shot correlation on test-3 compared to a non-domain-randomized baseline.
5. Interpretability: Attention maps derived from the cross-attention weights will be visualized overlayed on video frames. This provides qualitative insight into the model's spatial-temporal focus when inferring specific physical properties.

VII. Conclusion

This project proposes a systematic investigation into extracting quantitative physical knowledge from Video Foundation Models. By implementing a state-of-the-art probing methodology and introducing **domain randomization** as a targeted enhancement to address the critical Sim-to-Real challenge, this research aims to demonstrate the feasibility of robust, **zero-shot physical parameter estimation** from video. The expected outcomes include a novel, domain-randomized dataset, rigorous empirical validation of the proposed methodology's effectiveness in improving real-world generalization, and a functional prototype. Successful completion will contribute significant insights into the emergent physical reasoning capabilities of VFM^s and offer a practical advancement towards automating physically realistic digital content creation and enhancing robotic perception, potentially paving the way for more physically grounded AI systems capable of deeper interaction with the physical world.