PAVE: LEARNING PHYSICS-AWARE VISUAL REPRESENTATIONS FROM VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical knowledge is essential for physical reasoning, video generation, and robotic manipulation. However, general-purpose visual encoders primarily learn semantic and spatial features, with limited emphasis on physical properties and interactions. We introduce **PAVE**, a vision encoder that learns transferable physical representations from videos. For data, we curate PhysCentric, a corpus of 515,377 unlabeled video clips covering diverse physical interactions. For architecture, we combine sparsely routed experts with a shared expert to capture specialized and common physical cues. For training, we first learn physical dynamics through causal future-feature prediction, then apply Physics Knowledge Alignment using physical question-answer supervision and Baseline-Adjusted Cross-Entropy (BACE) to support physical reasoning. PAVE achieves 58.93% on MVBench and 70.04% on PhysBench, surpassing the strongest compared encoders by 3.08 and 2.01 percentage points, respectively. Downstream gains over the respective baselines include **9.7** percentage points in VideoPhy semantic adherence, **6.4** points in VSI-Bench average score, and **3.2** percentage points in average success rate across five RoboTwin 2.0 tasks, demonstrating its transferability. The data, code, and model will be publicly available.

1 INTRODUCTION

Encoding physical knowledge is essential for vision encoders that support physical reasoning, video generation, and embodied interaction (Xu et al., 2019; Zhang et al., 2026; Assran et al., 2025). Such knowledge includes properties such as mass and friction, as well as how objects respond to forces and contact. Some of these properties are difficult to infer from appearance alone but can be revealed through motion and interaction; for example, observing how a sliding object slows down provides clues about friction (Wu et al., 2015). However, widely used vision encoders, such as SigLIP 2 and DINOv3, primarily develop general-purpose semantic and spatial representations through image-text alignment and visual self-supervision (Tschannen et al., 2025; Siméoni et al., 2025) instead of focusing on physical knowledge. Image–text pretraining, exemplified by CLIP and SigLIP, provides broad semantic supervision but does not explicitly optimize physical prediction (Radford et al., 2021; Zhai et al., 2023). Although these representations can contain physical information, their training objectives do not explicitly prioritize the physical properties and interactions needed for downstream physical understanding. This motivates learning vision encoders that specifically target the extraction and representation of physical knowledge.

Several studies have explored physical representation learning by predicting interaction outcomes (Battaglia et al., 2016; Watters et al., 2017; Xu et al., 2019) or masked video features (Garrido et al., 2025; Assran et al., 2025). Despite this progress, two challenges remain. First, existing general-purpose video pretraining corpora may cover diverse objects and scenes while providing uneven coverage of physical interactions, such as sliding, collisions, and deformation. As a result, models pretrained on these corpora may capture visual semantics and temporal dynamics well, yet struggle to learn the underlying physical properties and interaction mechanisms. Second, predictive objectives optimize the reconstruction or prediction of visual features without explicitly identifying which information is useful for physical reasoning. These challenges motivate a physics-centric corpus that covers diverse physical interactions and a training paradigm that combines predictive pretraining with explicit supervision for physical reasoning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Overall framework of PAVE. (a)** PhysCentric comprises 515,377 clips from nine sources spanning real-world videos, robotic interactions, and physics simulations. **(b)** PAVE is trained through causal future-feature prediction followed by Physics Knowledge Alignment. **(c)** The learned representations improve downstream performance in video generation, spatial reasoning, and robotic manipulation.

To address these challenges, we develop a physics-centric data curation and training framework for building **PAVE**, a visual encoder that learns transferable physical representations from videos. As shown in Figure 1, for data, we curate **PhysCentric**, a corpus of 515,377 unlabeled clips that brings together real-world videos, robotic interactions, and controlled simulations covering diverse physical phenomena. For training, we begin with causal future-feature prediction to learn how physical scenes evolve, followed by Physics Knowledge Alignment that uses physical question-answer supervision to adapt the learned representations for downstream reasoning. During alignment, we introduce Baseline-Adjusted Cross-Entropy (BACE) to encourage PAVE to capture additional physical cues by improving correct-answer margins relative to a frozen baseline. To capture the distinct visual cues required by different physical phenomena, we equip PAVE with sparsely routed experts and a shared expert. Learned routing enables input-dependent specialization, while the shared expert captures common features across observations.

In our experiments, PAVE achieves 58.93% on MVBench (Li et al., 2024) and 70.04% on PhysBench (Chow et al., 2025), outperforming the strongest compared baselines by 3.08 and 2.01 percentage points, respectively. It also improves over the respective baselines by **9.7**, **6.4**, and **3.2** on video generation, spatial intelligence, and embodied AI, supporting the transferability of its physical representations. Ablations further support our data and training design. Our main contributions are:

- We curate **PhysCentric**, a corpus of over 515K clips unifying real-world videos, robotic interactions, and controlled simulations across diverse physical phenomena.
- We introduce **PAVE**, a visual encoder with a sparse MoE architecture that combines routed and shared experts to capture complementary physical cues.
- We develop a **two-stage** training framework combining predictive pretraining with Physics Knowledge Alignment, using answer supervision and BACE to adapt physical representations for downstream reasoning.
- We outperform the compared visual encoders across benchmarks and downstream tasks, demonstrating the effectiveness and transferability of our learned physical representations.

2 Related Work

2.1 Self-Supervised Vision Encoders

Self-supervised learning enables vision encoders to acquire transferable representations without manual annotations. One approach encourages consistency across augmented views, through contrastive learning in SimCLR (Chen et al., 2020), bootstrapping in BYOL (Grill et al., 2020), and self-distillation in DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2023). Another approach reconstructs masked visual content, as in MAE (He et al., 2022) for images and VideoMAE and VideoMAE V2 (Tong et al., 2022; Wang et al., 2023) for videos. Beyond pixel reconstruction, joint-embedding predictive approaches learn to predict target representations from visible context, progressing from images to large-scale video pretraining (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Recent work emphasizes spatial structure and motion through multi-layer token supervision (Mur-Labadia et al., 2026) and tracked trajectory reconstruction (Vandeghen et al., 2026). Complementary regularizers control feature variance and covariance in VICReg (Bardes et al., 2021) or encourage isotropic Gaussian features in LeJEPA (Balestriero & LeCun, 2025) to prevent collapse. These objectives do not explicitly prioritize physical properties and interactions; we instead target representations useful for physical understanding and reasoning.

2.2 Physics-Aware Visual Representation Learning

Physics-aware visual representation learning aims to encode physical properties and interactions from images and videos. Galileo infers object properties by combining visual observations with physical simulation (Wu et al., 2015), while interaction networks learn object-centric dynamics from states or video (Battaglia et al., 2016; Watters et al., 2017). Later work learns dense physical properties through interaction (Xu et al., 2019) and physical expectations through object-centric prediction (Piloto et al., 2022). Physion and CLEVRER evaluate physical prediction and causal reasoning (Bear et al., 2021; Yi et al., 2019), while PhysBench probes broader physical understanding in vision–language models (Chow et al., 2025). Recent studies investigate physical knowledge emerging from video pretraining and its organization within visual encoders (Garrido et al., 2025; Joseph et al., 2026), motivating transferable physics-aware representations.

A complementary line incorporates physical representations and supervision into generation and reasoning models. VideoPhy evaluates semantic adherence and physical commonsense in generated videos (Bansal et al., 2025). Feature alignment and generative preference feedback transfer or learn physical representations for video generation (Zhang et al., 2026; Ji et al., 2025). Physics-oriented instruction tuning targets physical commonsense and embodied reasoning (Azzolini et al., 2025; Lian et al., 2026), while transition-level supervision connects state changes to physical mechanisms (Wu et al., 2026b). Unlike approaches centered on a downstream system, we train an encoder for reuse across tasks. Architecturally, sparse expert routing expands capacity without activating every expert (Fedus et al., 2022) and has been extended to vision Transformers (Riquelme et al., 2021). Shared-expert designs separate common and routed computation (Dai et al., 2024); PAVE combines this principle with physics-oriented data and two-stage training.

3 PhysCentric: A Physics-Centric Video Dataset

Learning physical representations requires observing how objects move and respond to physical interactions. For example, videos of objects slowing down as they slide or deforming under pressure provide clues about friction and stiffness that are difficult to infer from a single image. These observations should also span diverse objects, environments, and viewpoints to support generalization. We therefore curate a large video corpus that brings together varied physical phenomena and visual settings for self-supervised pretraining.

To this end, we curate **PhysCentric**, a corpus of 515,377 unlabeled clips from 9 public sources, summarized in Table 1. The corpus combines observations from real-world videos, robot trajectories, and controlled or simulated scenes. Real-world videos capture egocentric activities and human–object interactions across varied environments and viewpoints. Robot trajectories record manipulation sequences and the resulting changes in object states, while controlled and simulated scenes

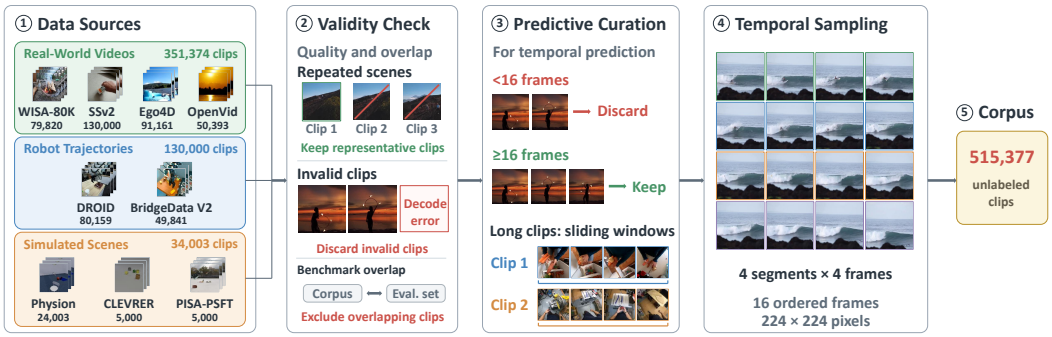

Figure 2: **Construction pipeline of PhysCentric.** Nine public datasets provide real-world videos, robot trajectories, and simulated scenes. The validity check removes redundant and invalid clips. Predictive curation discards clips with fewer than 16 frames and segments long videos using sliding windows to avoid excessive temporal gaps between sampled frames. Sampling four frames from each of four temporal segments yields 16 ordered frames at 224×224 resolution per clip, forming a corpus of 515,377 unlabeled clips.

Table 1: Composition of PhysCentric after source-specific curation and integrity checks.

Source	Clips	Source	Clips
WISA-80K	79,820	DROID	80,159
Ego4D	91,161	BridgeData V2	49,841
Something-Something V2	130,000	Physion	24,003
OpenVid	50,393	CLEVRER	5,000
PISA-PSFT	5,000		
Total clips			**515,377**

provide focused observations of selected physical processes. Together, these sources cover mechanical motion, collision and deformation, optical effects, thermal processes, and phase transitions.

Figure 2 summarizes the construction pipeline. We first convert each source dataset into a unified clip record. Short videos and frame sequences are treated as individual clips, long egocentric videos are segmented into shorter clips, and robot trajectories are sampled at the episode level to reduce redundancy. We then remove duplicate and invalid clips, exclude clips or source videos that overlap with downstream evaluation benchmarks, and retain only samples with at least 16 decodable frames. The 16-frame minimum and source-specific segmentation of long videos into shorter clips are intended to preserve sufficient temporal context for future-feature prediction while avoiding excessively large gaps between sampled frames. To prevent any single dataset from dominating the corpus, we further cap large or highly redundant source subsets. After extraction and filtering, we sample 16 frames from each retained clip to obtain a standardized model input. Specifically, each clip is divided into four equal temporal segments, from which four frames are sampled uniformly and arranged in temporal order. The resulting frames are resized to 224×224. Detailed extraction rules, filtering criteria, and sampling procedures are provided in Appendix A.

4 Method

4.1 Overview

Building on our physics-oriented video dataset, we aim to develop a vision encoder that learns physical properties and interactions from videos and makes this knowledge accessible to downstream models. Our framework consists of the PAVE encoder and two training stages. During predictive pretraining, PAVE learns temporal dynamics by predicting future latent features from an observed prefix while SIGReg (Balestriero & LeCun, 2025) prevents representation collapse. During physics knowledge alignment, physics question answering provides task-oriented supervision through a frozen language backbone, and BACE explicitly measures the improvement of the PAVE-augmented branch over a frozen baseline.

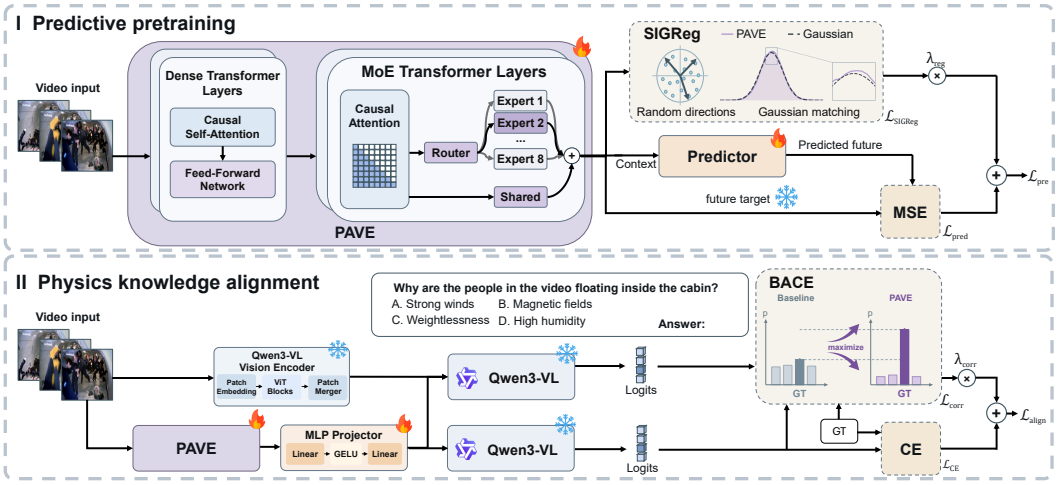

Figure 3: Training framework of PAVE. **(I) Predictive Pretraining.** PAVE combines causal dense and MoE layers, trained with future-feature prediction and SIGReg regularization. **(II) Physics Knowledge Alignment.** With the language backbone frozen, PAVE and its projector are optimized using answer cross-entropy and BACE to improve correct-answer margins relative to a frozen VLM.

4.2 PAVE Encoder

Physical properties are often not directly observable from appearance alone, but are revealed by how objects move, deform, and respond to interactions over time. Extracting such information therefore requires video representations that capture these evolving responses. PAVE is designed to learn these representations by modeling spatial interactions and temporal changes jointly. To capture dependencies across objects and time, PAVE uses a 12-block Transformer. Temporal causal attention is applied throughout to construct each time-step representation using only observations available up to that time.

Within this framework, different physical phenomena require sensitivity to different cues, such as motion changes during collisions and shape changes during deformation. To accommodate this variation, PAVE introduces sparse mixture-of-experts (MoE) modules in the final four blocks, following eight dense blocks that provide common feature processing. Each MoE module selects one routed expert for input-dependent processing and includes an always-active shared expert to retain a common pathway across inputs. Given a token representation $\mathbf{h}$ after causal attention, the outputs are combined as:

$$\mathrm{MoE}(\mathbf{h}) = E_{\mathrm{sh}}(\mathbf{h}) + \sum_{i=1}^{N} g_i(\mathbf{h}) E_i(\mathbf{h}), \tag{1}$$

where $N = 8$, E_{sh} and E_i denote the shared and routed experts, respectively, and $g_i(\mathbf{h})$ is nonzero only for the selected expert. Expert assignments are learned end to end, allowing specialization without predefined physical categories. In Appendix C, we analyze the physical information captured by each expert.

4.3 Stage I: Predictive Pretraining

To train PAVE to capture these evolving physical responses, we need both videos that reveal physical properties and an objective that encourages the encoder to learn from them. Our PhysCentric dataset provides the former. For the latter, we adopt future-feature prediction (Bardes et al., 2024): because predicting future features encourages the encoder to capture information relevant to these underlying physical factors.

Specifically, we split each 16-frame video V at a sampled temporal anchor into an observed prefix and a future segment, denoted by $\mathcal{M}_{\mathrm{ctx}}$ and $\mathcal{M}_{\mathrm{fut}}$, respectively. The context encoder extracts features from the observed prefix:

$$\mathbf{Z}_{\mathrm{ctx}} = f_\theta(V_{\mathcal{M}_{\mathrm{ctx}}}). \tag{2}$$

Given these context features and mask tokens at future positions, a predictor p_ϕ estimates the future features:

$$\hat{\mathbf{Z}}_{\text{tgt}} = p_\phi(\mathbf{Z}_{\text{ctx}}, \mathcal{M}_{\text{fut}}) . \tag{3}$$

To construct the training targets, we apply the same causal encoder to the complete video and select the features at future positions:

$$\mathbf{Z}_{\text{tgt}} = \text{sg}([f_\theta(V)]_{\mathcal{M}_{\text{fut}}}) , \tag{4}$$

where sg denotes stop-gradient. Causal attention ensures that each target feature incorporates only its current and preceding observations, preserving the available temporal context. The prediction branch receives only the observed prefix and learns to predict these future representations.

We supervise prediction with the mean squared error

$$\mathcal{L}_{\text{pred}} = \frac{1}{BMd} \left\| \hat{\mathbf{Z}}_{\text{tgt}} - \mathbf{Z}_{\text{tgt}} \right\|_F^2 , \tag{5}$$

where B is the batch size, M is the number of future target tokens, and d is the feature dimension. However, prediction alone admits a trivial solution in which the encoder produces identical features for different inputs. We therefore apply SIGReg (Balestriero & LeCun, 2025) to the context features. SIGReg projects the features along random directions and matches the resulting one-dimensional distributions to a standard Gaussian, discouraging representation collapse.

The overall pretraining loss is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{reg}} \mathcal{L}_{\text{SIGReg}}, \tag{6}$$

and we jointly optimize PAVE and the predictor.

4.4 Stage II: Physics Knowledge Alignment

Predictive pretraining encourages PAVE to capture information useful for anticipating future observations, but it may not emphasize the physical cues most relevant to downstream reasoning. We therefore introduce **Physics Knowledge Alignment**, which uses physics-related questions to identify the relevant physical information and encourage the model to capture this information in its representations. To encourage PAVE to contribute discriminative information beyond the frozen baseline, we introduce Baseline-Adjusted Cross-Entropy (BACE). BACE subtracts the scaled baseline logits from the PAVE-augmented option logits before applying cross-entropy. By encouraging PAVE to better distinguish the correct answer from the alternatives, this objective aims to improve its ability to extract the physical information needed to answer the question.

For each video–question pair (V, q), we obtain option logits from two branches sharing a frozen Qwen3-VL-2B language backbone (Bai et al., 2025):

$$\begin{aligned} \mathbf{z}_0 &= F_{\text{LM}}(E_0(V), q) , \\ \mathbf{z}_V &= F_{\text{LM}}([g_\psi(f_\theta(V)); E_0(V)], q) , \end{aligned} \tag{7}$$

where F_{LM} returns the option logits from the shared frozen language backbone, E_0 denotes the original frozen visual pathway, and f_θ and g_ψ denote PAVE and its projector, respectively. Both branches retain the original visual pathway. PAVE and its projector provide an additional visual pathway in the augmented branch, where $[\cdot;\cdot]$ denotes token concatenation. Only θ and ψ are updated during alignment. To guide PAVE toward features useful for physical question answering, we supervise the ground-truth answer with cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\log P_{\theta,\psi}(y_{\text{GT}} \mid V, q) , \tag{8}$$

where θ and ψ denote the parameters of PAVE and its projector, respectively.

To further encourage learning physical knowledge beyond the frozen baseline, BACE subtracts the scaled baseline option logits from the PAVE-augmented option logits and normalizes the resulting scores:

$$p_{\text{corr}}(k) = \frac{\exp(z_V(k) - \alpha \, \text{sg}(z_0(k)))}{\sum_{j=1}^{4} \exp(z_V(j) - \alpha \, \text{sg}(z_0(j)))}, \tag{9}$$

where z_V and z_0 are the augmented and baseline logits, α controls the correction strength, and sg denotes stop-gradient. When $\alpha = 1$, matching the baseline logits produces a uniform corrected

Table 2: Comparison of visual encoders evaluated on MVBench and PhysBench. Parameter counts indicate visual-encoder parameters activated per forward pass; baseline counts are approximate. † denotes the native visual encoder of Qwen3-VL-4B. The best and second-best results on each benchmark are shown in bold and underlined, respectively.

Visual encoder	Params (M)	MVBench ↑	PhysBench ↑
Qwen3-VL-4B† (Bai et al., 2025)	415	50.82%	55.86%
VideoMAE V2-L (Wang et al., 2023)	304	52.92%	<u>68.03%</u>
V-JEPA 2.1-L (Mur-Labadia et al., 2026)	305	53.83%	59.05%
SigLIP2-L (Tschannen et al., 2025)	316	51.88%	55.58%
DINOv3-L (Siméoni et al., 2025)	303	<u>55.85%</u>	60.90%
PAVE-Dense (Ours)	86	50.35%	59.08%
PAVE (Ours)	105	**58.93%**	**70.04%**

distribution. BACE therefore encourages PAVE to improve the correct option's relative score, with the aim of extracting physical information that supports this improvement. We combine the two supervision signals as:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}}, \qquad \mathcal{L}_{\text{corr}} = -\log p_{\text{corr}}(y_{\text{GT}}). \tag{10}$$

After alignment, we transfer the updated PAVE encoder to downstream tasks, including physical visual question answering and video generation and so on.

5 Experiments

5.1 Settings

Training Settings. We initialize PAVE from scratch and pretrain it on PhysCentric for 100 epochs. We then perform Physics Knowledge Alignment with Qwen3-VL-2B (Bai et al., 2025), updating PAVE and its projector while freezing the language model and its native visual pathway. More details are described in Appendix B. For pretraining ablations, we train all variants for 30 epochs due to computational resource constraints, while the model used for the main results uses the pretrained checkpoint at epoch 69.

Evaluation Protocol. All external encoders are evaluated through a shared Qwen3-VL-4B (Bai et al., 2025) interface. We freeze each encoder and train a new projector, the last four language-model blocks, and the final normalization layer on PhysBench for three epochs. The resulting checkpoint is evaluated on both PhysBench (Chow et al., 2025) and MVBench (Li et al., 2024). We report the accuracy and active visual encoder parameter counts. More implementation details are provided in Appendix B.

5.2 Main Results

Quantitative Results. We compare PAVE with representative image and video encoders under a unified evaluation protocol using Qwen3-VL-4B. As shown in Table 2, PAVE achieves 58.93% on MVBench and 70.04% on PhysBench, surpassing the strongest baselines, DINOv3-L and VideoMAE V2-L, by 3.08 and 2.01 percentage points, respectively, with approximately 65% fewer active parameters. These gains are consistent with PAVE's explicit focus on physical understanding: causal future prediction on PhysCentric encourages learning how objects move and interact, while Physics Knowledge Alignment adapts these representations for downstream reasoning, extending beyond the semantic and spatial features emphasized by encoders such as DINOv3. Compared with PAVE-Dense, PAVE improves accuracy by 8.58 and 10.96 percentage points, respectively, with active parameters increasing from 86M to 105M. This improvement suggests that input-dependent expert specialization helps accommodate diverse physical cues, such as motion changes during collisions, while the shared expert retains common features across inputs.

Table 3: Ablation of Physics Knowledge Alignment and BACE. Both alignment variants start from the same pretrained checkpoint and follow the same downstream evaluation protocol. Results are reported as accuracy (%).

Training stage	MVBench ↑	PhysBench ↑
Pretraining only	52.65%	64.85%
+ Alignment w/o BACE	56.43%	67.04%
+ Alignment w/ BACE	**58.93%**	**70.04%**

Table 4: Ablation on visual encoder initialization after 30 pretraining epochs.

Initialization	MVBench ↑	PhysBench ↑
VideoMAE V2-B	53.40%	59.60%
V-JEPA 2.1-B	50.20%	52.83%
Random (Ours)	**56.37%**	**61.40%**

Table 5: Ablation on attention and masking. Random masking uses non-causal attention.

Strategy	MVBench ↑	PhysBench ↑
Bidirectional	52.92%	54.55%
Random mask	55.07%	59.78%
Causal (Ours)	**56.37%**	**61.40%**

5.3 Ablation Study

Ablation on Training Stage. We isolate the contributions of Physics Knowledge Alignment and BACE using the same pretrained checkpoint and downstream evaluation protocol. Without BACE, alignment updates the entire PAVE encoder and its temporary Qwen3-VL-2B projector using standard cross-entropy. With BACE, alignment additionally incorporates detached logits from the frozen Qwen-only baseline into the baseline-adjusted objective described in Section 4.4. As shown in Table 3, Physics Knowledge Alignment improves performance on both benchmarks, and adding BACE yields further gains of 2.50 and 3.00 percentage points on MVBench and PhysBench, respectively. These results suggest that Physics Knowledge Alignment makes pretrained physical representations more useful for downstream reasoning by emphasizing cues relevant to answering physical questions. BACE further encourages the encoder to capture discriminative information beyond the frozen baseline by rewarding improvements in the correct answer's margin over competing options.

Ablation on Initialization. We compare initialization strategies using the same PAVE architecture and a fixed budget of 30 pretraining epochs. For pretrained initialization, each source Transformer block initializes the corresponding PAVE block; within each sparse block, the source FFN weights are copied to all routed experts and the shared expert. As shown in Table 4, random initialization achieves the best performance, surpassing VideoMAE V2-B initialization by 2.97 percentage points on MVBench and 1.80 percentage points on PhysBench. These results suggest that, in this setting, training from scratch on our physics-oriented corpus and predictive objective is more effective than adapting representations initialized from pretrained dense encoders.

Ablation on Causal Mask. We compare causal future prediction with bidirectional attention and non-causal random masking after 30 pretraining epochs. As shown in Table 5, causal prediction outperforms bidirectional attention by 3.45 and 6.85 percentage points on MVBench and PhysBench, respectively, and random masking by 1.30 and 1.62 points. These gains suggest that predicting unseen future states encourages learning motion trends and interaction-driven changes, such as how contact alters an object's trajectory, while non-causal reconstruction may rely more on local correlations between surrounding frames.

5.4 Downstream Evaluation

Video Generation. To evaluate the transferability of PAVE's physical representations to video generation, we adopt the VideoREPA framework (Zhang et al., 2026) and construct two CogVideoX-2B (Yang et al., 2025b) variants that differ only in the teacher encoder: VideoMAE (Wang et al., 2023) for the baseline and PAVE for our method, while keeping all other training settings identical between the two variants.

We evaluate both variants on a 310-video subset of VideoPhy (Bansal et al., 2025) and report Semantic Adherence (SA) and Physical Commonsense (PC). The training and evaluation details are described in Appendix B. As shown in Table 6, PAVE+VideoREPA achieves an average SA score of 0.5581 and a PC score of 0.2839, improving upon VideoMAE+VideoREPA by 0.0968 and 0.0452, respectively. PAVE also consistently improves performance across all three interaction categories. In particular, on Fluid–Fluid interactions, the PC score increases from 0.3404 to 0.4468. These results support the transferability of PAVE's physical representations to video generation, with particularly clear gains in physical consistency.

Table 6: Video generation results on a 310-video subset of VideoPhy.

Teacher Encoder	**SA↑**	**PC↑**	**Solid–Solid SA↑ / PC↑**	**Solid–Fluid SA↑ / PC↑**	**Fluid–Fluid SA↑ / PC↑**
VideoMAE (Wang et al., 2023)	0.4613	0.2387	0.3462 / 0.1231	0.6015 / 0.3158	0.3830 / 0.3404
PAVE (Ours)	**0.5581**	**0.2839**	**0.4308 / 0.1539**	**0.7143 / 0.3534**	**0.4681 / 0.4468**

Spatial Intelligence. We assess the contribution of PAVE to visual spatial reasoning by comparing the supervised fine-tuned GUIDE (Wang et al., 2026a) model with its PAVE-enhanced variant under the same VSI-Bench evaluation protocol. GUIDE+PAVE adds an additional branch to GUIDE and integrates the resulting geometric features into the language model through cross-attention, while keeping the remaining model and evaluation settings unchanged. The architectural and training details are provided in Appendix B. As reported in Table 7, GUIDE+PAVE achieves an average score of 69.8 on VSI-Bench (Yang et al., 2025a), improving over GUIDE by 6.4 percentage points. The improvement is consistent across all eight categories, showing that PAVE is particularly effective for spatial relation modeling and geometric reasoning.

Table 7: Spatial reasoning results on VSI-Bench (%).

Models	**Avg.↑**	**Numerical Answer**				**Multiple-Choice Answer**			
		Obj. Count↑	**Abs. Dist.↑**	**Obj. Size↑**	**Room Size↑**	**Rel. Dis.↑**	**Rel. Dir.↑**	**Route Plan↑**	**Appr. Order↑**
Spatial-MLLM-4B (Wu et al., 2026a)	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VG-LLM-4B (Zheng et al., 2026)	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4
GUIDE (Wang et al., 2026a)	63.4	69.4	48.8	71.3	64.2	65.4	80.2	37.1	71.2
GUIDE+PAVE	**69.8**	**69.6**	**56.6**	**75.0**	**73.5**	**71.4**	**87.3**	**47.4**	**77.7**

Embodied AI. To evaluate the transferability of PAVE's physical representations to robotic manipulation, we conduct simulation experiments on five tasks from RoboTwin 2.0 (Chen et al., 2025), as summarized in Table 8. The training and evaluation details are described in Appendix B. Success rates are computed according to the official RoboTwin task-completion criteria. As shown in Table 8, PAVE achieves an average success rate of 69.4%, exceeding LingBot-VA (Li et al., 2026) and VideoMAE by 6.2 and 3.2 percentage points, respectively. PAVE achieves the highest success rates on three of the five tasks: Hanging Mug (30%), Place Bread Skillet (81%), and Place Burger Fries (91%), improving upon VideoMAE by 13, 5, and 1 percentage points, respectively. Although the gains vary across tasks, the improvement in average success rate supports the transferability of PAVE's physical representations to robotic manipulation.

Table 8: Success rates on five RoboTwin 2.0 tasks.

Method	Hanging Mug↑	Place Can Basket↑	Place Bread Skillet↑	Place Burger Fries↑	Stack Bowls Three↑	Average↑
LingBot-VA (Li et al., 2026)	19%	69%	72%	82%	74%	63.2%
+ VideoMAE (Wang et al., 2023)	17%	**70%**	76%	90%	**78%**	66.2%
+ PAVE (Ours)	**30%**	69%	**81%**	**91%**	76%	**69.4%**

6 Conclusion

We presented PAVE, a visual encoder that learns transferable physical representations from videos. PAVE combines pretraining on the PhysCentric corpus with causal future prediction, sparse expert specialization, and Physics Knowledge Alignment to capture physical dynamics and adapt them for downstream reasoning. Experiments show that PAVE outperforms the strongest compared encoders on MVBench and PhysBench while using substantially fewer active parameters. Further evaluations demonstrate its transferability to video generation and robotic manipulation. We hope our work inspires further research on representation learning.

AI USE STATEMENT

We used GPT-5.6 to assist with drafting and refining the manuscript text, including improving its grammar, clarity, and readability. All AI-assisted content was reviewed and revised by the authors to ensure accuracy and consistency with the intended technical meaning. The authors take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Our method is described in Section 4. Details of the datasets, preprocessing procedures, evaluation protocols, and experimental configurations are provided in Appendix A and Appendix B to support the reproducibility of the reported results.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 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. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, volume 2025, pp. 97959–98108, 2025.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5843–5851. IEEE, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Physmaster: Mastering physical representation for video generation via reinforcement learning. *arXiv preprint arXiv:2510.13809*, 2025.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. *arXiv preprint arXiv:2602.07050*, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

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.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Shijie Lian, Bin Yu, Xiaopeng Lin, Changti Wu, Hang Yuan, Xiaolin Hu, Zhaolong Shen, Yuzhuo Miao, Haishan Liu, Yuxuan Tian, Yukun Shi, Cong Huang, and Kai Chen. Physbrain 1.0 technical report, 2026. URL https://arxiv.org/abs/2605.15298.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-jepa 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision*, pp. 671–689. Springer, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pp. 1045–1064, 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.

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9): 1257–1267, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in neural information processing systems*, 34:8583–8595, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Renaud Vandeghen, Fida Mohammad Thoker, Marc Van Droogenbroeck, and Bernard Ghanem. Trackmae: Video representation learning via track mask and predict. *arXiv preprint arXiv:2603.27268*, 2026.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on robot learning*, pp. 1723–1736. PMLR, 2023.

Chongyu Wang, Ting Huang, Chunyu Sun, Xinyu Ning, Di Wang, and Hao Tang. Let geometry guide: Layer-wise unrolling of geometric priors in multimodal llms. *arXiv preprint arXiv:2604.05695*, 2026a.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. *Advances in Neural Information Processing Systems*, 38:5388–5416, 2026b.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560. IEEE, 2023.

Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30, 2017.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026a.

Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 28, 2015.

Shang Wu, Haoran Lu, Songling Liu, Chenwei Xu, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, Zhaoran Wang, and Han Liu. Apt: Atomic physical transitions for causal video-language understanding. *arXiv preprint arXiv:2606.18586*, 2026b.

Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. *arXiv preprint arXiv:1906.03853*, 2019.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025a.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 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. *Advances in Neural Information Processing Systems*, 38:122647–122676, 2026.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026.

A Additional Dataset Details

Source-aware curation. The source datasets differ substantially in storage format and temporal granularity: some provide short videos, some provide frame directories, and robot datasets organize observations as long episodes. We first map every source into a common clip record while retaining its dataset name, source identifier, temporal location, frame rate, resolution, and decoding backend. Short videos from WISA-80K (Wang et al., 2026b), Physion (Bear et al., 2021), CLEVRER (Yi et al., 2019), Something-Something V2 (Goyal et al., 2017), and BridgeData V2 (Walke et al., 2023) are treated as individual clips, whereas Ego4D (Grauman et al., 2022) videos are divided into short egocentric clips. For DROID (Khazatsky et al., 2024), we retain at most four temporally ordered clips from each trajectory so that long robot episodes do not dominate the corpus. For OpenVid (Nan et al., 2025), repeated source identifiers are removed, only one scene is retained per original video, and the highly repetitive CelebV subset is capped at 4,000 clips. We additionally cap large sources, including Something-Something V2 at 130,000 clips, to preserve diversity across domains. We also include the PSFT training subset of PISA (Li et al., 2025).

Integrity filtering. After source-specific conversion, we remove missing records, duplicated source entries, videos that cannot be decoded, and clips with fewer than 16 decodable frames. We then scan the generated corpus and retain a sample only when all of its processed frames and metadata are present. This procedure avoids relying on potentially inaccurate container-level frame counts and makes every retained sample traceable to its source.

Unified temporal and spatial processing. Given a clip with T decodable frames, we divide its timeline into four equal temporal segments and uniformly sample four frames from each segment. This deterministic $4{\times}4$ strategy covers the complete clip while preserving short-range temporal detail within each segment. The selected RGB frames are decoded in temporal order, resized to $224{\times}224$, and stored as JPEG frames. The accompanying metadata record the exact source-frame indices, original frame rate, duration, spatial resolution, and source path. During pre-training, the same spatial transformation is applied consistently to all frames in a clip. We use multi-scale cropping with scales $\{1, 0.875, 0.75, 0.66\}$, resize the crop to 224×224, and normalize RGB values using ImageNet statistics.

B More Training and Evaluation Details

Training Details. PAVE contains 12 Transformer blocks, with dense FFNs in the first eight blocks and eight routed experts with top-1 routing plus one shared expert in each of the last four blocks. We pretrain the encoder and predictor on all 515,377 PhysCentric clips using AdamW (Loshchilov & Hutter, 2017) and bfloat16 precision on eight GPUs. The batch size is 16 per GPU and 128 globally. The learning rate increases linearly from 1×10^{-6} to 7.5×10^{-5} over the first three epochs, then follows a cosine schedule to 2.5×10^{-5}. We use weight decay 0.05, gradient clipping at 1.0, and a SIGReg (Balestriero & LeCun, 2025) loss weight of 0.3. Pretraining runs for 100 epochs; the final model's alignment stage starts from the epoch-69 checkpoint.

Physics Knowledge Alignment uses over 30,000 video question-answering (VQA) examples. We update the complete PAVE encoder and its projector while keeping Qwen3-VL-2B's language model and native visual pathway frozen (Bai et al., 2025). We use AdamW with constant learning rates of 2×10^{-5} for PAVE and 1×10^{-6} for the projector, a global batch size of 16, and bfloat16 precision. The BACE correction coefficient and loss weight are both set to 1.0. Frozen baseline option logits are computed once and cached offline. The alignment projector is discarded before downstream transfer.

Evaluation Protocol on PhysBench and MVBench. Each external encoder is frozen and connected to Qwen3-VL-4B (Bai et al., 2025) through a newly initialized projector that maps its temporal features into the language-model embedding space. The native visual pathway is retained. We train the projector, the last four language-model blocks, and the final normalization layer on the PhysBench (Chow et al., 2025) training split for three epochs. AdamW uses learning rates of 1×10^{-4} for the projector and 1×10^{-5} for the trainable language-model parameters, with a global batch size of 16, weight decay 0.01, and bfloat16 precision. PAVE receives 16 frames at 224×224

with temporal stride 4; baseline encoders use the same temporal sampling and their native spatial preprocessing.

We evaluate the same checkpoint on the 6,002-example PhysBench test set and the 4,000-example MVBench (Li et al., 2024) test set, without further tuning. Both benchmarks are formulated as multiple-choice video question answering, and performance is measured by exact-match top-1 accuracy. No MVBench training data are used. Parameter counts refer to visual-encoder parameters activated in one forward pass, which equal total parameter counts for dense encoders.

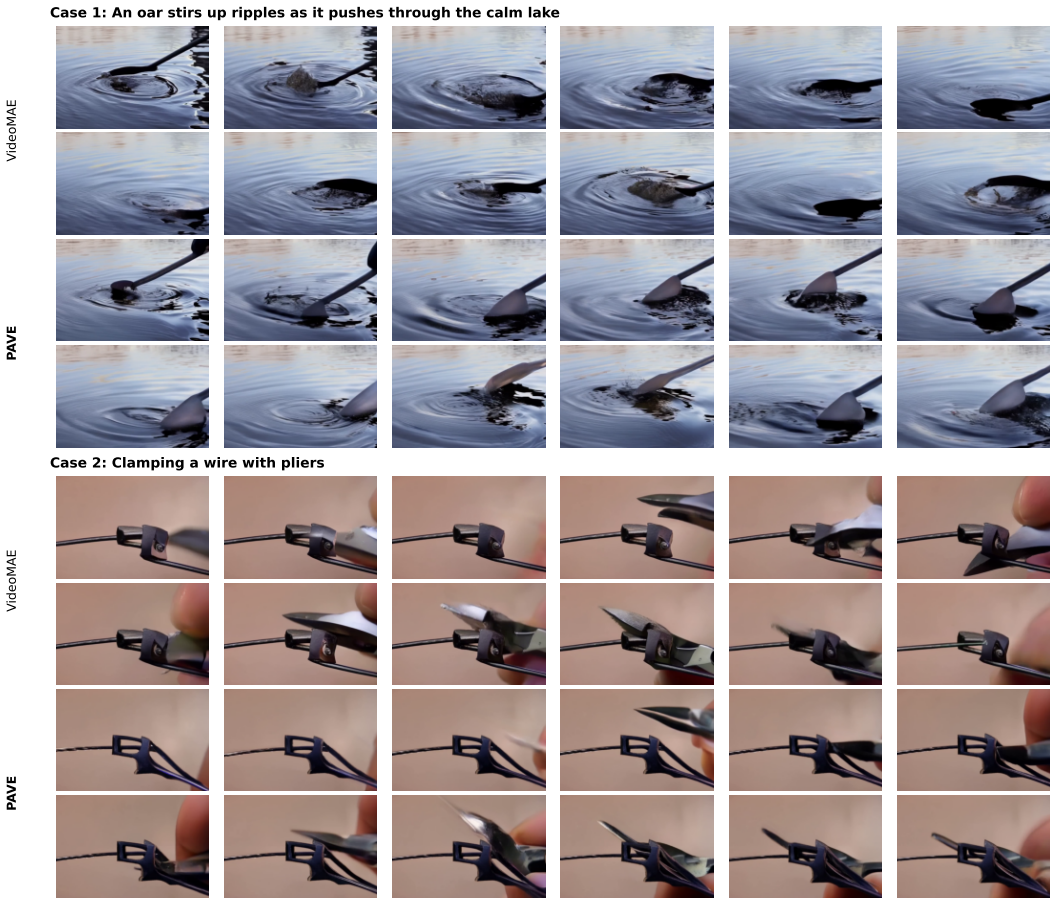

Figure 4: Qualitative comparison between VideoMAE and PAVE under the VideoREPA framework on representative VideoPhy examples. For each case, the two methods are visualized at matched temporal positions to facilitate comparison of the generated physical dynamics over time.

Evaluation Protocol on Video Generation We evaluate the transferability of PAVE to video generation using the VideoREPA framework (Zhang et al., 2026) with CogVideoX-2B (Yang et al., 2025b) as the generation backbone. Specifically, we construct two variants using VideoMAE and PAVE as teacher encoders, respectively. For PAVE+VideoREPA, we replace the original VideoMAE teacher with the PAVE encoder while keeping the student architecture, training objective, optimization schedule, data configuration, and video resolution unchanged. Both variants are trained on the same 3.2K-video subset of WISAt (Wang et al., 2026b) using an identical training configuration with a batch size of 4, a learning rate of 2×10^{-6}, 1000 optimization steps, and bf16 precision.

We evaluate both models on a 310-video subset of VideoPhy (Bansal et al., 2025). Following the VideoPhy evaluation protocol, we report Semantic Adherence (SA) and Physical Commonsense (PC), together with category-wise results for Solid–Solid, Solid–Fluid, and Fluid–Fluid interactions. We additionally provide qualitative comparisons on representative examples in Figure 4. For each example, frames from VideoMAE and PAVE are sampled at identical normalized temporal positions, enabling direct comparison of their generated physical dynamics over time.



Figure 5: **Embodied manipulation on Hanging Mug and Place Bread Skillet.** For each task, the three rows show LingBot-VA, LingBot-VA with VideoMAE, and LingBot-VA with PAVE, respectively. Columns show the progression of each rollout.

Evaluation Protocol on Spatial Intelligence. GUIDE+PAVE builds on the supervised fine-tuned GUIDE model (Wang et al., 2026a), which uses Qwen3-VL-4B as its base vision-language model (Bai et al., 2025). We augment GUIDE with PAVE, an additional physics-aware visual encoder that extracts physical representations from input videos.

For each input video, we sample 16 frames and organize them into four groups of four frames. PAVE features are temporally pooled into physical tokens of shape $[B, 8, 768]$. These tokens are integrated into the language model through cross-attention at six injection points, allowing the model to progressively incorporate physical cues for spatial reasoning without directly concatenating the physical tokens with the input sequence.

We train GUIDE+PAVE in two stages, keeping the PAVE encoder frozen throughout. In the first stage, we freeze the language model and optimize only the connection layers between PAVE and the language model to align their representations. In the second stage, we freeze the connection layers and further optimize the language model. Both GUIDE and GUIDE+PAVE are evaluated after supervised fine-tuning under the same evaluation protocol.

Evaluation Protocol on Embodied AI. We train jointly on the five RoboTwin 2.0 (Chen et al., 2025) tasks using 550 demonstrations per task. Each observation includes one head-camera view and two wrist-camera views. Starting from LingBot-VA (Li et al., 2026), we construct two variants by incorporating video feature tokens from VideoMAE or PAVE into the frozen embodied MoT backbone via cross-attention. Both variants use the same policy architecture, feature integration scheme, and training configuration. We evaluate each method over 100 episodes per task, using identical environment seeds across methods and instructions unseen during training.

Figures 5 and 6 compare LingBot-VA, its VideoMAE-augmented variant, and its PAVE-augmented variant on four RoboTwin 2.0 tasks. Frames progress from left to right. PAVE's rollouts show sustained grasps during mug placement, coordinated placement of food items, and stable bowl stacking.

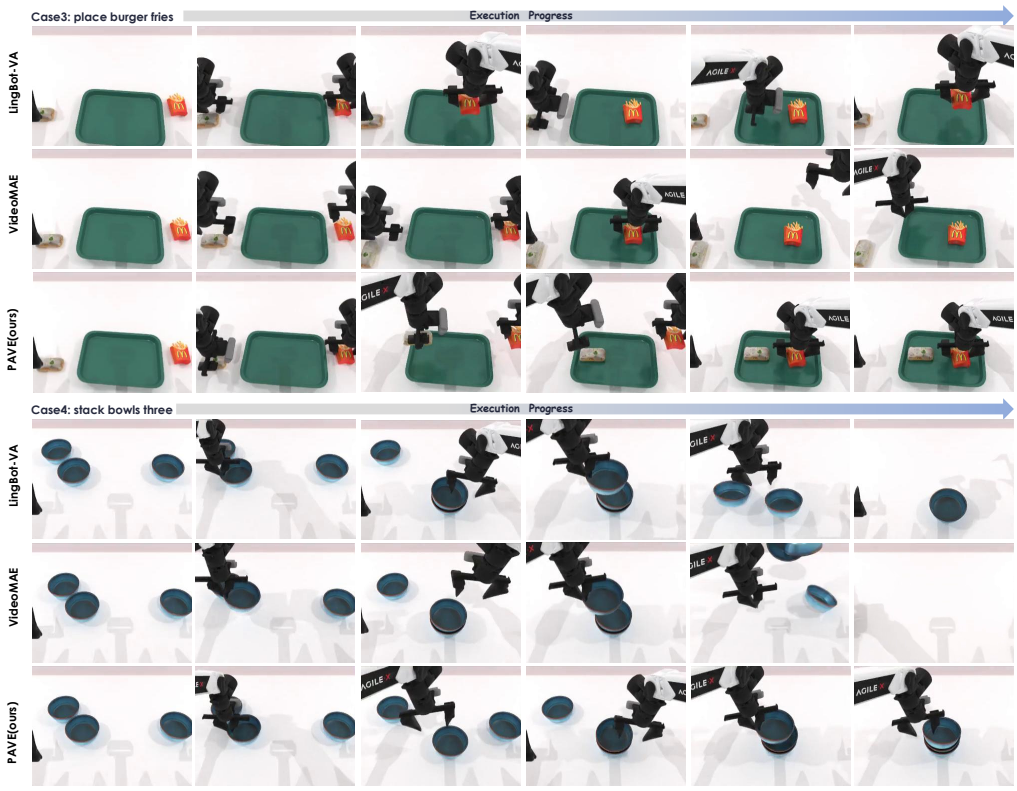

Figure 6: **Embodied manipulation on Place Burger Fries and Stack Bowls Three.** The row ordering follows Figure 5. PAVE's rollouts show placement of both food items and completion of the three-bowl stack.

Table 9: Ablation on PhysCentric data scale. Accuracy (%).

Data scale	**MVBench ↑**	**PhysBench ↑**
20% data	51.60%	52.70%
Full data	**56.37%**	**61.40%**

These examples complement the quantitative results by illustrating how the transferred physical representations support object manipulation over time.

C More Analysis

Ablation on Data Composition. To examine performance in a reduced-data regime, we construct a source-stratified subset by retaining one fifth of the clips from every PhysCentric source. This preserves the relative source mixture while reducing the pre-training corpus to 20% of its original size while keeping its data composition unchanged. Both models are evaluated with the same downstream protocol. As shown in Table 9, reducing the corpus to 20% lowers accuracy from 56.37% to 51.60% on MVBench and from 61.40% to 52.70% on PhysBench. The larger degradation on PhysBench suggests that broad coverage of physical events is particularly important for learning representations that support physical reasoning.

Expert Routing Visualization. Consistent with the MoE design in Section 4.2, PAVE exhibits distinct routing preferences across different physical scenes (Figure 7). In the illustrated examples, Expert 3 receives 79.5%–95.9% of routing assignments during hand and gripper manipulation, while Expert 6 receives 79.6%–89.6% in scenes involving rigid-body motion and contact. Expert 7 dom-

inates natural scenes featuring water and clouds, with routing shares of 72.7%–100.0%. These recurring preferences across different videos suggest that routed experts specialize in complementary visual cues associated with different physical interactions. Such input-dependent allocation supports PAVE's division of processing: routed experts provide differentiated computation for heterogeneous physical scenes, complementing the common feature processing of the shared expert.

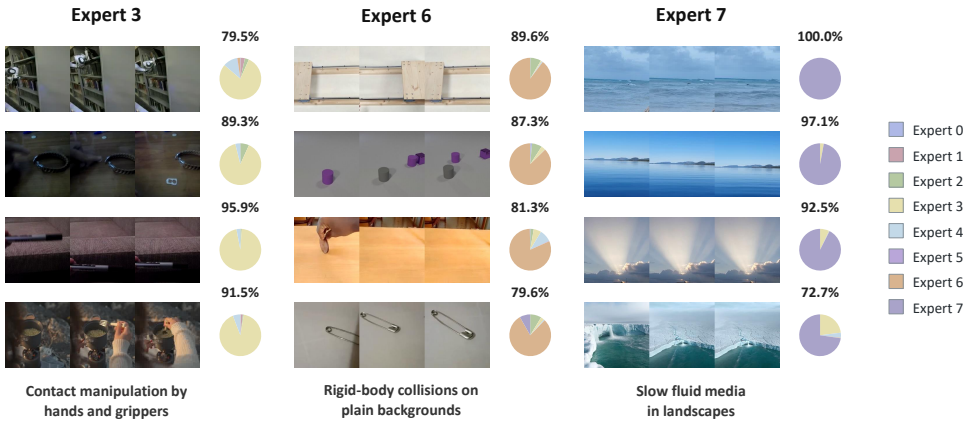

Figure 7: **Expert routing on representative physical scenes.** Columns show hand and gripper manipulation (Expert 3), rigid-body motion and contact (Expert 6), and natural scenes with water and clouds (Expert 7). Each example contains three video frames and a routing distribution over eight experts; percentages indicate the dominant expert's routing share.

Latent Visualization. Figure 8 compares PAVE with VideoMAE V2 on solar-eclipse and flower-wilting videos. In these examples, PAVE's feature maps more clearly delineate the evolving occlusion boundary and changing petal shapes while separating these structures from the background. These structures provide observable cues to physical state, reflecting changes in occlusion geometry during the eclipse and object deformation during wilting. Preserving such cues is consistent with our causal future-feature prediction objective, which encourages representations to retain state information useful for anticipating subsequent changes. These visualizations provide qualitative support for PAVE's sensitivity to evolving physical states.

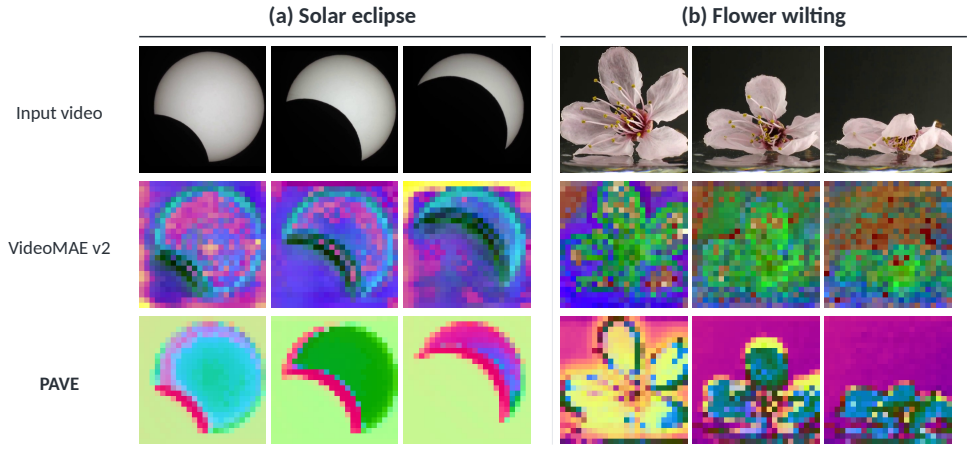

Figure 8: **Latent feature visualization.** Input frames (top) and corresponding feature maps from VideoMAE V2 (middle) and PAVE (bottom) for a solar eclipse and flower wilting. Columns within each example follow temporal order.