Next-Frame Prediction as a Reliability-Aware Training Paradigm for Robust Vision Encoders

Anonymous submission

Abstract

Foundation models deployed in dynamic domains like robotics and autonomous systems suffer from critical reliability failures, including temporal inconsistencies and vulnerability to sensor noise, stemming from their training on static, disconnected images. To bridge this reliability gap, we propose a lightweight, reliability-aware training paradigm that distills temporal knowledge from video into a standard single-image encoder. By training a predictor to estimate the feature representation of a future frame, our method implicitly forces the backbone model to learn real-world dynamics, enhancing robustness to transient visual artifacts and promoting temporally stable representations. This self-supervised objective instills geometric and physical priors without relying on brittle external modules like optical flow estimators. Remarkably, when pre-trained on only a single, 2-hour uncurated video, our method sets a new state-of-the-art for DINO-style video distillation on downstream tasks like detection and segmentation, which we use as quantifiable proxies for robust scene understanding. Our work presents a practical and efficient approach for improving the trustworthiness and dependable performance of vision encoders for safe deployment in operational settings.

Introduction and Related Work

The deployment of foundation models in critical domains is hampered because their "stochastic nature and sensitivity to context make them vulnerable to distribution shifts, sensor noise, and hallucinations," limiting their safe deployment. This is especially true for the emerging field of Physical AI (Yang et al. 2025; Edge AI Foundation 2025), where systems must interact with the real world. A core source of unreliability stems from pre-training vision encoders on vast datasets of static, independent images (Yin et al. 2023). This paradigm fails to capture the temporal coherence of physical reality, leading to critical failure modes: representations lack temporal consistency, causing perceptual "jitter," and are brittle to the continuous domain shifts in sensor streams. For an embodied agent, an unstable object representation is a potential catalyst for catastrophic physical failure, making model trustworthiness paramount.

Self-supervised learning (SSL) has produced powerful encoders from static images. Seminal approaches used instance discrimination, either contrastively (e.g., MoCo, SimCLR; He et al. 2020; Chen et al. 2020) or non-contrastively (e.g., BYOL, DINO; Grill et al. 2020; Caron et al. 2021), to learn invariant representations. More recently, masked image modeling (MIM) has become dominant, using a BERT-like objective to reconstruct masked patches (e.g., BEiT, MAE; Bao et al. 2022; He et al. 2022). Despite their success, these methods learn from a "shuffled deck" of images, rendering them blind to the arrow of time and the causal structure of the world.

To overcome this, video SSL aims to learn from temporal data. One approach is to train dedicated, often computationally intensive, video backbones that explicitly model space-time features (Bertasius, Wang, and Torresani 2021; Tong et al. 2022). While powerful, their high inference cost

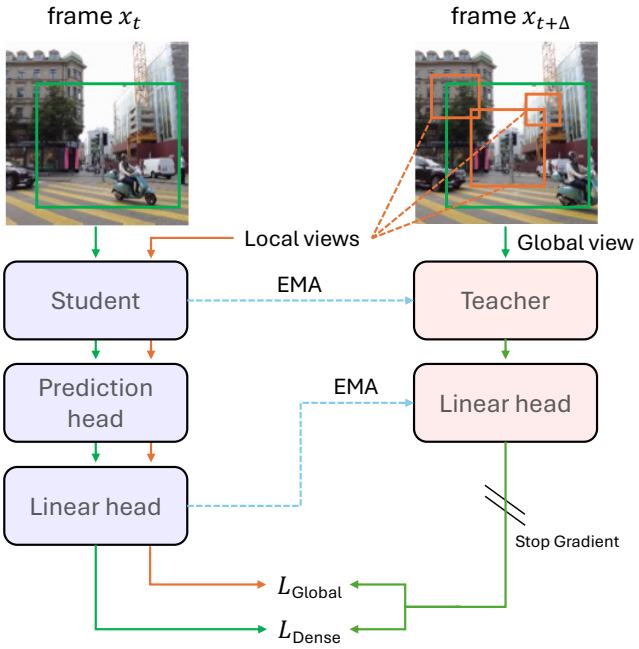

Figure 1: **Overview of our reliability-aware training.** The student encodes a frame and a dense prediction head is used to estimate the features of the teacher for the next frame. This prediction task encourages temporally robust representations. Exponential moving average (EMA) is used to update the teacher.

makes them ill-suited for many latency-sensitive robotics applications. A more pragmatic approach, which we follow, is to distill knowledge from videos into a fast *single-image* encoder. The idea of learning by predicting future representations is itself an established concept (Han et al. 2020). Recent methods have applied this to distillation: DoRA (Venkataramanan et al. 2024) uses object tracking to guide the process, but still optimizes frames independently. PooDLe (Wang et al. 2024) enforces temporal equivariance using optical flow, but this introduces a brittle dependency. Optical flow estimators are a known source of unreliability, often failing under the exact sensor degradations (e.g., motion blur, occlusions) that reliable systems must handle, making it a poor foundation for a trustworthy model.

Our approach. We propose a reliability-aware training paradigm that avoids these pitfalls. A student encoder learns to regress the teacher's dense, patch-level feature map of a future frame from the current one. This forces the model to learn an implicit understanding of local scene dynamics and transformations, rather than just predicting a single, global context vector. This simple objective forces the encoder to internalize an implicit model of motion and causality, embedding temporal and geometric priors directly into its weights. The lightweight prediction head is discarded after training, yielding a standard, fast single-image encoder that is inherently more robust and produces temporally dependable representations. This is especially critical for Vision–Language–Action (VLA) models (Brohan et al. 2025; Kim et al. 2024) that currently rely on static-image encoders, inheriting their reliability vulnerabilities, which can cause cascading failures in the downstream action-generation policy.

Contributions.

- **A Reliability-Aware Training Paradigm:** We introduce a next-frame prediction objective that instills temporal robustness and geometric consistency into an off-the-shelf image encoder without inference-time overhead.
- **A Metric for Dependable Performance:** Our approach improves feature consistency over time, a crucial characteristic for dependable performance in robotics that is not captured by static accuracy metrics alone. We demonstrate this via strong performance on dense prediction tasks, which serve as a proxy for robust scene understanding.

Proposed Method

Problem setup. Given an unlabeled video $V = \{x_1, \dots, x_T\}$ we form clips $\mathcal{C}_t = \{x_{t+i\Delta}\}_{i=0}^{K-1}$, where Δ is a stride hyper-parameter (default Δ=30 frames) and K=3. We apply a pre-crop and then mostly follow (Caron et al. 2021) for each frame, but apply the same pre-crop and global crop to all frames to allow dense prediction. Local views are obtained for each frame starting with $x_{t+\Delta}$.

Architecture. Our architecture (Figure 1) follows the self-distillation framework of DINO (Caron et al. 2021). Both teacher and student share a ViT-S backbone and projection head. In contrast to DINO, we insert a lightweight prediction head, comprising a 2-layer MLP and two attention blocks, between the student's backbone and its projection head. This head is exclusive to the student and is discarded after training, ensuring no inference-time overhead. Teacher weights are an exponential moving average (EMA) of the student's, with no gradients flowing through the teacher.

Loss functions. The teacher processes global views from the second frame onward, while the student processes all global views except the last, plus all local views. We optimize two complementary losses.

(i) Dense next-frame loss. This is the core of our reliability-aware objective. For every consecutive pair of frames $(x_j, x_{j+\Delta})$, the student S must predict the teacher T's patch tokens of the future frame, $\hat{z}^{\mathrm{S}}_{j\to j+\Delta}$, using only the current frame x_j. The loss is:

$$\mathcal{L}_{\text{dense}} = \frac{1}{K-1} \sum_{j=t}^{t+(K-2)\Delta} \text{P-CE}\big(\sigma_{\tau_{\mathrm{T}}}(z^{\mathrm{T}}_{j+\Delta}), \sigma_{\tau_{\mathrm{S}}}(\hat{z}^{\mathrm{S}}_{j\to j+\Delta})\big), \tag{1}$$

where $z^{\mathrm{T}}_{j+\Delta}$ are the teacher's patch tokens, P-CE is the per-patch averaged cross-entropy, and σ_τ is softmax with temperature τ.

(ii) Global loss. To maintain feature quality, we supplement with a standard intra-frame consistency loss. For each future frame x_j and its L local crops, we enforce consistency between the teacher's global [CLS] token output $\bar{z}^{\mathrm{T}}_j$ and the student's local view predictions $\tilde{z}^{\mathrm{S}}_{j,\ell}$.

$$\mathcal{L}_{\text{global}} = \frac{1}{(K-1)L} \sum_{j=t+\Delta}^{t+(K-1)\Delta} \sum_{\ell=1}^{L} \text{CE}\big(\sigma_{\tau_{\mathrm{T}}}(\bar{z}^{\mathrm{T}}_j), \sigma_{\tau_{\mathrm{S}}}(\tilde{z}^{\mathrm{S}}_{j,\ell})\big). \tag{2}$$

Total objective. The final training loss is the unweighted average $\mathcal{L} = 0.5\,\mathcal{L}_{\text{dense}} + 0.5\,\mathcal{L}_{\text{global}}$. The dense loss operates on *patch tokens* across time, while the global loss uses *[CLS] tokens* within a single frame.

Why Prediction Fosters Reliability. Regressing future features forces the student to learn an implicit model of real-world dynamics, directly addressing several key failure modes highlighted by the workshop:

- **Robustness to Sensor Degradations:** Predicting across time encourages the model to learn a representation that acts like a low-pass filter over transient sensor phenomena. It must learn to ignore high-frequency noise like specular highlights, single-frame motion blur, or compression artifacts, as these have no causal bearing on the future state of the underlying scene.
- **Mitigating Temporal Hallucinations:** The objective provides a strong inductive bias towards a form of object permanence, penalizing temporally inconsistent representations. This encourages smooth and predictable embeddings, preventing erratic behavior in downstream control policies that rely on stable object tracking.
- **Learning Physical Priors:** To successfully predict future features in a dynamic scene, the model is incentivized to implicitly disentangle ego-motion from independent object motion. This forces it to learn fundamental priors about

Table 1: Semantic segmentation on ADE20K (mIoU). Best results are bold, second-best are underlined. † Result from (Wang et al. 2024).

Method	Backbone	UperNet	Fast-LP
PooDLe†	ResNet-50	36.6	14.6
All following methods use a ViT-S/16 backbone			
iBOT	ViT-S/16	**38.0**	**19.7**
Ours (dense+global)	ViT-S/16	<u>36.4</u>	<u>18.3</u>
DoRA	ViT-S/16	35.0	17.0
Analysis of DINO Baselines			
DINO (frames only)	ViT-S/16	31.7	12.9
+ pre-crop	ViT-S/16	35.1	15.8
+ time-aug (Δ=5)	ViT-S/16	34.9	15.9
Ablation of Our Method			
dense-loss only	ViT-S/16	36.2	17.4
global-loss only	ViT-S/16	34.5	15.6

3D consistency and the plausible motion of objects in the world.

- **Handling Distribution Shifts:** Self-supervised training on video from an operational environment allows the model to learn the specific motion patterns and appearance distributions of that domain, enhancing robustness against shifts from static, out-of-distribution image datasets.

Experiments

We evaluate our approach using semantic segmentation, object detection, and instance segmentation as downstream tasks. We use these tasks as quantifiable proxies for model reliability; success reflects a model's ability to form a robust and accurate understanding of its environment, which is critical for dependable performance. Our central hypothesis is that learning to model real-world dynamics forces the encoder to develop more robust and geometrically consistent features, and that this enhanced robustness transfers directly to improved performance on challenging static scene understanding tasks. Instance segmentation, in particular, serves as a demanding test of pixel-level precision, a key requirement for safe robotic interaction.

Experimental Setup

Data and Training. All models are trained from scratch on the public *Walking Tours Venice* 60 fps video (2 hours) (Venkataramanan et al. 2024). We intentionally use this limited and constrained data setting to highlight the data efficiency of our approach and its ability to learn robust priors from readily available, in-domain video streams without requiring massive, general-purpose datasets. The video provides dense scenes crowded with people in an unconstrained environment, which is ideal for learning object interatctions, occlusions, etc. Clips of $K{=}3$ frames are sampled with a stride of $\Delta{=}30$ (0.5s) unless stated otherwise. We use a ViT-S/16 backbone and train on four NVIDIA RTX 4090 GPUs for 100 epochs. Teacher weights are used for all evaluations. Training takes about one day.

Table 2: Object detection on MS-COCO (mAP). Best results are bold, second-best are underlined.

Method	Pre-train Data	Box mAP
Pre-trained on ImageNet-1K		
iBOT	Img-1K	**49.4**
AttMask	Img-1K	48.7
DINO	Img-1K	48.4
Pre-trained on WT Venice (2 hours)		
iBOT	WT Venice	**42.2**
Ours	WT Venice	<u>41.6</u>
DoRA	WT Venice	41.0
DINO + pre-crop	WT Venice	40.3

Table 3: Instance segmentation on MS-COCO (mAP). Best results are bold, second-best are underlined.

Method	Pre-train Data	Mask mAP
Pre-trained on WT Venice (2 hours)		
iBOT	WT Venice	**38.5**
Ours	WT Venice	<u>38.1</u>
DoRA	WT Venice	37.6
DINO + pre-crop	WT Venice	37.1

Evaluation Protocol. We evaluate semantic segmentation on ADE20K (Zhou et al. 2017) using UperNet. Object detection and instance segmentation are evaluated on MS-COCO 2017 (Lin et al. 2014) using a Mask R-CNN head. We follow the standard iBOT evaluation protocol (Zhou, Yang, and Loy 2022) for comparability with prior work. All baselines are retrained by us, except for the official DoRA checkpoint.

Main Results

Dense Scene Understanding. Table 1 shows our results for semantic segmentation. We first note that iBOT, which belongs to the more complex family of masked image modeling (MIM) methods, sets the overall performance ceiling. However, **among DINO-style self-distillation approaches, our method is the clear top performer.** Our full model achieves 36.4 mIoU, significantly outperforming other video-distillation methods like DoRA (+1.4 points) and all DINO baselines. The ablation study confirms this success stems from our predictive objective: using our 'dense-loss only' achieves 36.2 mIoU, retaining nearly all gains, while simpler 'time-augmentation' provides no benefit. This demonstrates that our reliability-aware prediction task, not just temporal proximity, is the key to advancing the state-of-the-art for this class of models.

Object-Level Reliability. As shown in Tables 2 and 3, our method's reliability extends to object-centric tasks on COCO. For object detection, it surpasses DoRA and a strong DINO baseline. More importantly, it achieves competitive performance on instance segmentation, a much harder task requiring pixel-perfect delineation. This strong result in predicting precise object masks is compelling evidence that our

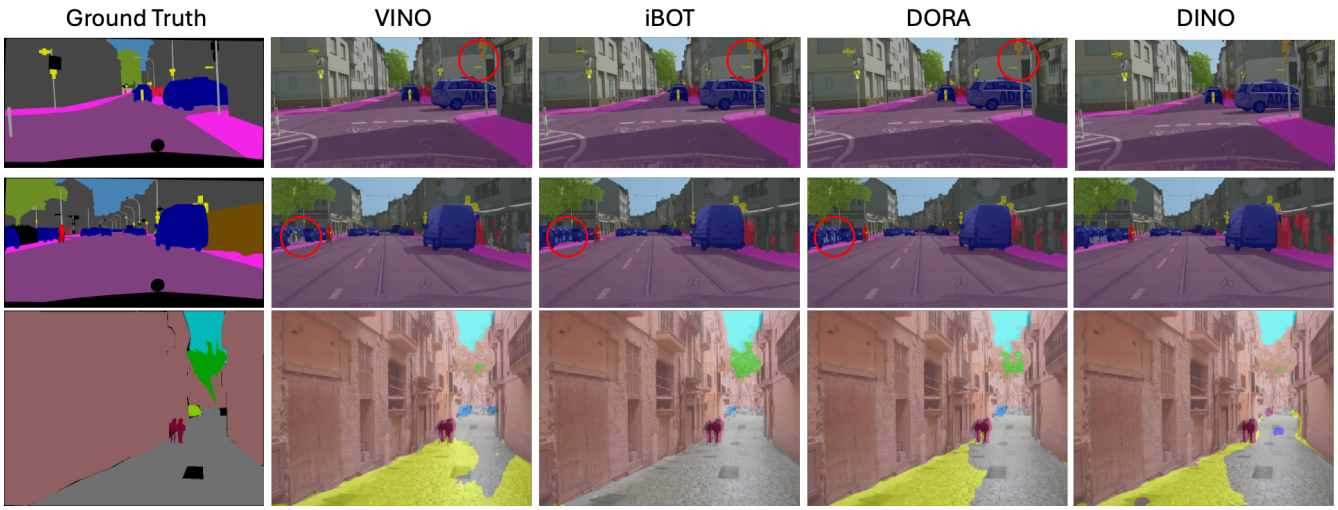

Figure 2: Qualitative comparison of semantic segmentation. Red circles highlight key areas where our method exhibits enhanced reliability. Note the reduction in segmentation artifacts and improved boundary definition within the highlighted regions compared to baselines. These subtle but critical improvements showcase a more robust scene understanding essential for dependable perception. Top two pictures are from Cityscapes, bottom from ADE20k.

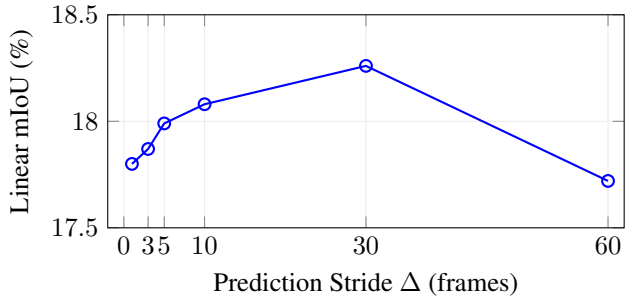

Figure 3: Analyzing the temporal horizon for robust feature learning. Performance peaks at a stride of Δ=30 frames (0.5s), suggesting an optimal timescale for learning meaningful dynamics.

reliability-aware training produces spatially accurate and dependable features, a non-negotiable property for safe robotic manipulation. This is achieved without the failure modes of external trackers or flow estimators.

Visualizing Reliability Improvements. Figure 2 visualizes our model's performance, using red circles to highlight key regions where methods differ. In these areas, our approach mitigates common failure modes by producing sharper, more geometrically plausible object boundaries and reducing noisy, spurious pixel classifications (e.g., on walls or roads). This improved fine-grained detail is direct evidence of a more dependable scene representation, which is critical for safety-critical tasks requiring precise spatial awareness.

Impact of Prediction Horizon. We analyze the impact of the prediction horizon in Figure 3. Reliability, measured by linear probe mIoU, steadily improves as the stride Δ increases, peaking at 30 frames (0.5s). The slight decline at 60 frames suggests that beyond a certain point, the prediction task becomes too difficult, introducing noise into the training signal. This confirms that learning to predict over a meaningful but finite temporal gap is key for building robust representations that are dependable for time-critical tasks.

Limitations and Future Work. Our work demonstrates the effectiveness of temporal prediction on a single, high-quality video. A natural limitation is the narrow diversity of scenes and motion patterns. Future work should focus on scaling this approach to large, diverse video datasets such as Ego4D or Something-Something-V2 to learn more universal models of real-world dynamics. Furthermore, we plan to explore more sophisticated predictive architectures and integrate our reliability-aware encoder directly into a Vision-Language-Action model to quantify the downstream impact on robotic task success rates and failure mode reduction in physical environments.

Conclusion

The safe deployment of foundation models in operational settings is contingent on overcoming their fundamental unreliability in dynamic environments. In this work, we introduced a lightweight, reliability-aware training method that distills temporal priors from video into a standard single-image encoder. By training the model to predict future representations, we directly encourage robustness to sensor noise and mitigate temporal inconsistencies. Our experiments, spanning semantic and instance segmentation as well as object detection, show that this simple principle is an efficient and powerful mechanism for learning dependable representations, setting a new performance standard for DINO-style video distillation. This approach is a practical step towards building the trustworthy and robust foundation models necessary for the safe deployment of *Physical AI* in the real world.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *Int. Conf. Learn. Represent.*

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Int. Conf. Mach. Learn.*

Brohan, A.; et al. 2025. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734*.

Caron, M.; Touvron, H.; Misra, I.; J'egou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Int. Conf. Comput. Vis.*

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Int. Conf. Mach. Learn.*, 1597–1607.

Edge AI Foundation. 2025. The Robots Are Coming—Physical AI and the Edge Opportunity. urlhttps://www.edgeaifoundation.org/edgeai-content/the-robots-are-coming-physical-ai-and-the-edge-opportunity/. Accessed 6 May 2025.

Grill, J.-B.; Strub, F.; Altch'e, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Pires, B.; Guo, Z. D.; Azar, M.; Piot, B.; Guez, A.; Pietquin, O.; Kavukcuoglu, K.; Larochelle, H.; Lanctot, M.; and Schmitt, S. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *Adv. Neural Inform. Process. Syst.*

Han, T.; Xiong, B.; Zolfaghari, M.; Morariu, V. I.; and Davis, L. S. 2020. Representation Forecasting for Anticipating Future Video Representations. In *Eur. Conf. Comput. Vis.*

He, K.; Chen, X.; Xie, S.; Li, Y.; Doll'ar, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9729–9738.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; et al. 2024. OpenVLA: An Open-Source Vision–Language–Action Model. *arXiv preprint arXiv:2406.09246*.

Lin, T.-Y.; Maire, M.; Belongie, S.; and et al. 2014. Microsoft COCO: Common Objects in Context. *Eur. Conf. Comput. Vis.*

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Adv. Neural Inform. Process. Syst.*

Venkataramanan, S.; Rizve, M. N.; Carreira, J.; Asano, Y. M.; and Avrithis, Y. 2024. Is ImageNet Worth 1 Video? Learning Strong Image Encoders from 1 Long Unlabelled Video. In *Int. Conf. Learn. Represent.*

Wang, A. N.; Hoang, C.; Xiong, Y.; LeCun, Y.; and Ren, M. 2024. PooDLe: Pooled and Dense Self-Supervised Learning from Naturalistic Videos. *Int. Conf. Learn. Represent.*

Yang, X.; et al. 2025. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

Zhou, B.; Zhao, H.; Puig, X.; and et al. 2017. Scene Parsing through ADE20K Dataset. *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhou, K.; Yang, J.; and Loy, C. C. 2022. iBOT: Image BERT Pre-Training With Online Tokenizer. In *Int. Conf. Learn. Represent.*