You Don't Need Strong Assumptions: Visual Representation Learning via Temporal Differences

Ninad Daithankar[*1], **Alexi Gladstone**[*1], **Yann LeCun**[2], **Heng Ji**[1]
[1]UIUC [2]New York University
temporal-difference-vision.github.io github.com/ninaddaithankar/TDV

Abstract

Progress in AI has largely been driven by methods that assume less. As compute and data increase, approaches with weaker inductive biases generally outperform those with stronger assumptions. This is particularly characteristic of the field of Visual Representation Learning, where approaches have gone from being dominated by Supervised Learning, to Weakly Supervised Learning, to the now widespread success of Self-Supervised Learning without human labels. Yet, even modern Self-Supervised Learning approaches still depend on strong inductive biases such as augmentations, masking, or cropping. If this trend holds, even these remaining biases should become bottlenecks at scale—and our experiments confirm this: the optimal strength of inductive biases decreases as data grows. This motivates the search for approaches that rely on fewer assumptions. To this end, we introduce **Temporal Difference in Vision** (**TDV**), a new paradigm for self-supervised learning from video that avoids existing inductive biases, relying instead on a causal assumption that the past causes the future. TDV functions by jointly training an image encoder and a motion encoder so that the current frame's representation plus the encoded motion equals the next frame's representation. Despite not leveraging any strong inductive biases, TDV matches state-of-the-art recipes on dense spatial tasks, laying the foundation for representation learning without strong assumptions.

1 Introduction

Deep learning has achieved remarkable progress over the last decade and a half, advancing from simple object classification [1] to high-resolution image generation [2] and sophisticated cross-modal reasoning [3]. This progress has largely been driven by methods that more effectively leverage increasing data and computation [4, 5], where approaches with weaker *inductive biases*[1] tend to outperform those with stronger assumptions as scale increases [6–12].

This principle is illustrated by the evolution of visual representation learning, where progress has largely been driven by approaches using progressively weaker inductive biases. For example, early supervised learning with convolutional neural networks (CNNs) [1, 13] assumed that human-annotated labels captured the semantic structure of images, while convolutional architectures imposed spatial locality biases on representations. Moving away from labels, self-supervised contrastive approaches such as SimCLR [14] and MoCo [15] instead pulled augmented views of the same image together and pushed different images apart, but made strong assumptions about the distances between negative pairs. To address this flaw, self-distillation approaches [16] relaxed these assumptions via a slow-moving teacher, and the subsequent adoption of Vision Transformers (ViTs) [7, 17] discarded the

[*]Equal Contribution. Correspondence to Alexi Gladstone: alexigladstone@gmail.com. Work done while supported as a Flapping Airplanes Fellow.

[1]We broadly use the term inductive biases and assumptions interchangeably in this work.

Preprint.

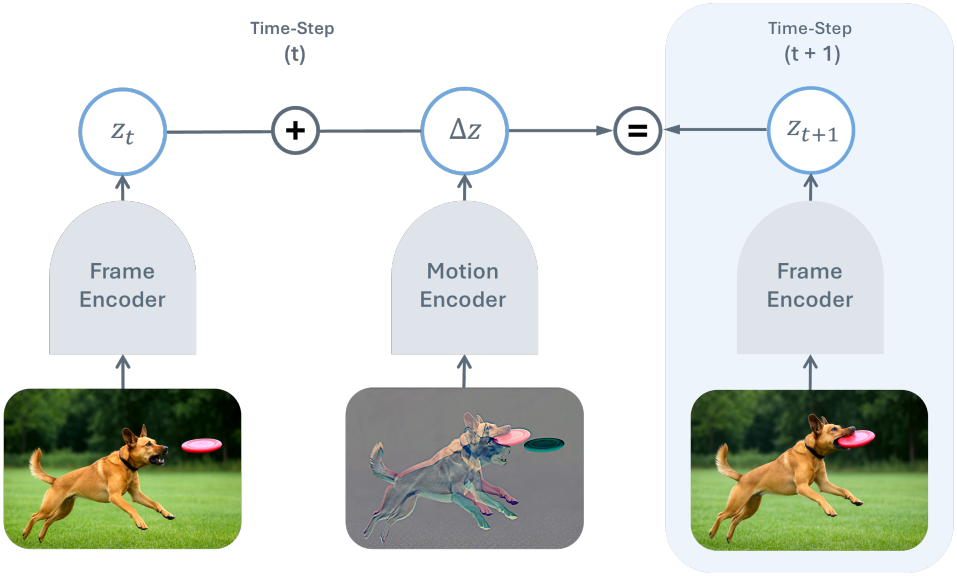

Figure 1: **TDV Frame and Motion Encoding Intuition.** TDV learns to encode frames such that the current frame's representation, when added to a learned motion encoding, predicts the next frame's representation. Because video has high temporal consistency, the raw RGB pixel difference between frames is intrinsically lower rank than the frames themselves, shown here as the edge outlines of a dog and a frisbee. The motion encoder compresses these high-dimensional RGB differences into abstract motion-level features.

locality and translation equivariance biases of CNNs in favor of global attention. Now, modern approaches combining self-distillation with ViTs achieve state-of-the-art performance [8, 9].

Interestingly, this pattern of weaker assumptions leading to better performance mirrors biological evolution, where innate instincts play a role analogous to hardcoded inductive biases. Across animals, more capable species tend to hardcode less behavior into their genome—insects rely heavily on innate behavioral programs, while mammals depend substantially more on learned behavior [18, 19]. This trend is even more distinct in primates, and most pronounced in humans, who rely heavily on learning from experience rather than hardcoded behavior [20, 21]. In both the case of visual representation learning as well as hardcoded behavior for biological intelligence, less hardcoded structure enables greater asymptotic performance given sufficient scale.

To further support this principle, we empirically test how the optimal strength of inductive biases changes with data scale (Figure 3). We find that as data scale increases, weaker inductive biases outperform stronger ones asymptotically—reinforcing that minimizing assumptions becomes increasingly important as scale increases.

Motivated by this trend, we argue for a new approach to visual representation learning that avoids the inductive biases relied upon by existing methods (we discuss these biases in Section A). Removing them naively, however, leaves no learning signal and collapses the representation (Table 1).

Therefore, a natural question emerges—*"What assumptions should our model have, if not reliant on existing inductive biases?"* We argue for assuming *causality*: that causes precede their effects, and the immediate future is therefore predictable from the past. This principle is foundational across physics, from classical mechanics to relativistic field theories [22] to modern formulations of quantum theory [23].[2] Unlike existing inductive biases for representation learning, we argue causality is weak, and domain agnostic. Additionally, because causality is inherently temporal, applying it points towards training over video, rather than images.[3]

[2]We do not claim the stronger thesis of determinism—that the past *uniquely* determines the future—which is incompatible with quantum mechanics under standard assumptions [24]. We assume only the weaker principle, that the immediate future is generally predictable from the past, sufficient to provide a learning signal.

[3]Learning image encoders from video departs from common practice in representation learning, which historically trains them on image datasets.

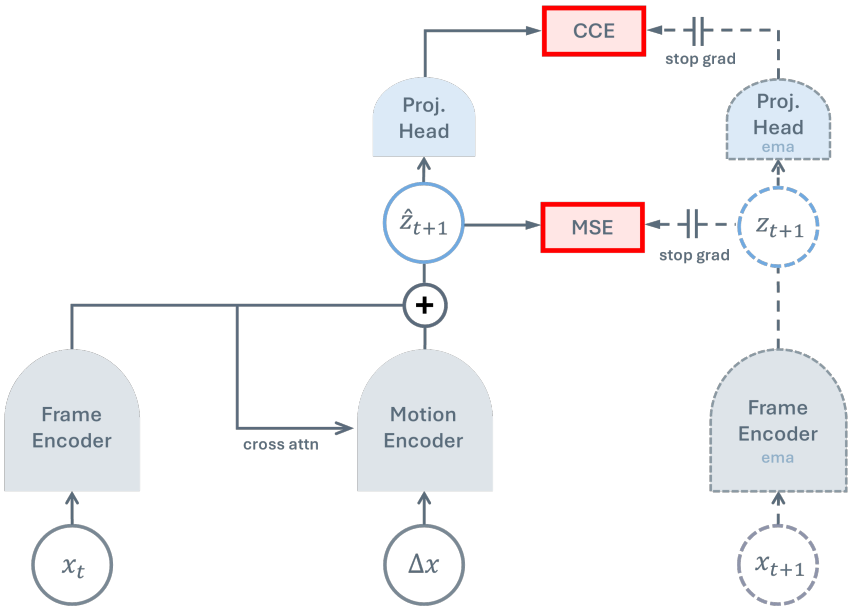

Figure 2: **TDV Architecture.** TDV predicts the next frame's representation by adding a learned motion vector to the current frame's representation. *Left (student):* the frame encoder embeds the current frame, while the motion encoder turns the raw pixel difference between frames into a latent motion shift, conditioned on the current frame via cross-attention. Their sum is the predicted representation of the next frame. *Right (teacher):* an EMA copy of the frame encoder embeds the true next frame to supply the target. Two losses act on the prediction: a mean-squared error on the representations enforces the causal next-frame constraint, and a DINO-style [17] cross-entropy on the projection heads prevents collapse. Stop-gradients block the teacher from receiving gradients. Figure style inspired by [25, 26].

To achieve this, we jointly train a frame encoder and a motion encoder so that, given two consecutive frames, the embedding of the current frame plus the embedding of the frame delta matches the embedding of the next frame (visualized in Figure 1). Because consecutive frames are close in time, and video has high temporal consistency, the frame delta is intrinsically low-rank, encouraging the motion encoder to capture compact spatial change rather than full scene appearance. Given its similarity to Temporal Difference in Reinforcement Learning [27], we call our approach **Temporal Difference in Vision (TDV)**. TDV naturally enables learning without restrictive inductive biases or modality-dependent assumptions. Empirical results demonstrate TDV is able to learn dense spatial features comparable to state-of-the-art visual encoders such as DINO [17] and iBOT [28] without relying on strong assumptions during pretraining.

Our contributions are as follows:

- We confirm our hypothesis regarding the importance of weaker assumptions as scale increases through controlled experiments, further motivating TDV.
- We present TDV, a new paradigm for learning visual representations that avoids the strong inductive biases of existing approaches.
- We demonstrate promising empirical performance for TDV, achieving dense spatial features on par with modern approaches that leverage stronger inductive biases.

2 Related Work and Background

2.1 Self-Supervised Representation Learning

Self-supervised representation learning has the goal of learning representations without any labels. Early work in this domain learned representations primarily via autoencoding [29, 30], where models

were trained to reconstruct inputs directly in pixel space. Over time, the field has shifted primarily from raw pixel space reconstruction towards Joint Embedding Predictive Architectures (JEPAs) [31], where prediction is done in a latent space as opposed to in the raw pixel space. Such models abstract away irrelevant, unpredictable information, such as background pixels in a scene, in favor of modeling more important information—a form of learning with weaker inductive biases. Recent empirical [9, 26] and theoretical [32, 33] evidence reinforces the benefits of JEPAs, where raw pixel space reconstruction is often theoretically predicted and empirically observed to produce less informative features for perception [32].

The progression within the JEPA family illustrates a recurring pattern: methods with weaker inductive biases have steadily displaced those with stronger ones. Early JEPA approaches relied on contrastive objectives [14, 15], which prevent collapse by pushing apart representations of distinct images while pulling together augmented views of the same image. However, contrastive methods impose a strong relational prior—that randomly sampled images should be dissimilar in representation space—which is only approximately correct, since sampled pairs frequently depict semantically related content. They also depend on large batches of negative samples, limiting scalability [34]. Self-distillation approaches [16, 17] relax this prior entirely by replacing negative pairs with a slow-moving teacher network: the student is trained to match the teacher's output on a different view of the same image, while the teacher is updated as an exponential moving average of the student. This eliminates the negative relational assumption between images, with centering and stop-gradient mechanisms preventing trivial collapse. Modern state-of-the-art vision foundation models such as DINOV3 [9] and V-JEPA2 [26] build on this self-distillation paradigm,[4] reflecting the broader trajectory of the field toward progressively weaker inductive biases.

2.2 Temporal Difference for Representation Learning

Despite this trajectory of weakening inductive biases, modern self-distillation approaches still rely on image-level inductive biases such as cropping, masking, or augmentations. A natural alternative is to source paired views from time itself, using temporally adjacent frames in a video. Several works explore this direction [35–39]. Feng et al. [40] train supervised models over temporal difference features, minimizing mutual information to disentangle task-relevant motion from noise. Wang et al. [41] model low-level frame deltas for action recognition, but rely on a global channel attention mechanism to recalibrate features across long-range differences. Maes et al. [42] predict future frames in latent space, but target world modeling rather than transferable representations. Most closely related to TDV, Midway Networks [43] learn representations directly from temporal differences in video, adding an invariance objective over cropped patches to target semantic performance. In each case, the temporal signal is paired with an additional inductive bias—supervision, attention recalibration, a world-modeling objective, or augmentation-based invariance. With TDV, we instead focus on learning from temporal difference alone, without any such biases.

3 TDV Approach

3.1 TDV Intuition

Learning representations without strong inductive biases is notoriously challenging [14, 44, 45]; removing assumptions such as augmentations or masking often leads to degraded representations or collapse. We confirm this by removing key inductive biases in the well-known DINO [17] recipe, observing poor performance and eventual collapse (Table 1). These results, and our deeper motivation to remove inductive biases, raise a natural question: *"What is the weakest assumption that still provides sufficient signal to avoid collapse?"*

Answering this requires understanding why assumptions hurt performance in the first place. We argue that assumptions encode beliefs that are only approximately correct [46, 47], and at scale these approximations restrict what can be learned [6, 7, 10]. Instead, an assumption that is exactly, rather than approximately, correct, would impose no such bottleneck, providing a learning signal without restricting what can ultimately be learned.[5]

[4] Some of these architectures, such as DINOV3, are technically considered Joint-Embedding Architecture (JEA) variants, due to not conditioning on a latent variable z.

[5] By "exactly correct" we refer to a property of the learning objective, not a claim about causality itself: next-frame prediction imposes no invariance constraint, and therefore never requires the encoder to discard a factor of variation. Augmentation- and

We argue such an assumption exists: *causality*—the principle that the past is predictive of the future. This principle is foundational across physics [22, 23], with classical mechanics serving as a canonical example—an object's position, velocity, and acceleration are sufficient to predict its trajectory. Unlike assumptions such as "augmented views should be invariant" or "masked and unmasked images should be similar," causality is domain-agnostic and, we argue, exactly rather than approximately correct. Perhaps this assumption is part of the reason autoregressive Large-Language Models have been so successful [3, 48, 49], as they assume causality, which reflects the data-generating procedure of language.

Table 1: **Removing DINO's Inductive Biases Degrades Performance.** Progressively removing DINO's augmentations for pre-training on SSV2 degrades KNN performance, and eventually causes representation collapse. TDV, by contrast, avoids collapse without these inductive biases. 2G and 8L denote 2 random global and 8 random local crops respectively. Full augmentations includes random flip, color jitter, Gaussian blur, and solarization.

Setup	↑Top-1	↑Top-5	Avoids Collapse
2G + 8L, full aug	24.63	40.19	✓
- augmentations	16.44	26.82	✓
- local crops	13.64	23.07	✓
- random crop on 1G	8.04	14.91	✓
- random crop on both G	0.84	2.37	×
Full TDV Recipe (ours)	8.79	17.05	✓

Leveraging causality, however, is non-trivial. Image representations are traditionally learned from static image datasets [8, 17, 44], which lack the temporal dimension causality requires.[6] We therefore argue for learning image representations from *video*, where consecutive frames provide the temporal structure causality demands.

Specifically, we train an image encoder jointly with a motion encoder such that the image encoder's representation of a frame, added to the motion encoder's representation of the change between frames, yields the image encoder's representation of the next frame (visualized in Figure 1). Intuitively, this motion representation generally has low intrinsic rank, since the semantic change between consecutive frames is typically small. By analogy to Temporal Difference in Reinforcement Learning [27], we call this approach **Temporal Difference in Vision (TDV)** (Figure 2).

Beyond its motivation from causality, TDV can also be viewed as a form of self-distillation [8, 16, 51]. However, rather than forcing invariance across hand-crafted augmentations such as cropping, rotating, or masking, the "augmentation" is induced by time, with temporally consecutive frames serving as the two views. The changes produced by this temporal augmentation are then modeled explicitly by the motion encoder, rather than being discarded via an invariance objective. This can be viewed as a learned, latent-space analog of the motion vectors used in classical video codecs [52], which similarly represent video as a frame plus the motion to the next. Intuitively, this objective forces TDV's representations to be sufficiently informative of the current frame as well as rich enough to predict the next frame.

3.2 TDV Architecture

Having established causality as our guiding assumption, we now derive the architecture for TDV. Our goal is to follow a simple principle: *the representation of a frame, combined with the change that occurs between frames, should yield the representation of the next frame.*

Learning a representation space. Following our principle, we first need a way to map the frames from RGB space into a meaningful representation space. We therefore learn a **frame encoder** f_θ that maps each frame x_t to a sequence of token embeddings:

$$z_t = f_\theta(x_t) \in \mathbb{R}^{n \times D}, \tag{1}$$

where z_t is the representation for frame x_t, n is the number of spatial patches plus an additional `[CLS]` token, and D is the embedding dimension. Our causal principle then becomes a constraint in this embedding space: the change between frames, when encoded appropriately, should be sufficient to predict the next frame's embedding.

masking-based objectives instead enforce invariance to a chosen transformation, discarding the corresponding information by construction, which can degrade downstream performance on tasks where that information is needed (Section A).

[6] One could apply causality within a single image by predicting one patch from another [50]; however, this does not reflect causality as it operates in the world, where causes precede effects in *time*.

Encoding change in representation space. The raw RGB difference $\Delta x_t = x_{t+1} - x_t$ captures what changed in pixel space, but we need to map this into a corresponding shift Δz_t in the latent space. Importantly, Δx_t is intrinsically lower rank than the frames themselves, as the background scene pixels remain largely unchanged between adjacent frames, and only moving regions contribute a non-zero signal (as visualized in Figure 1). We therefore learn a **motion encoder** m_ϕ that takes the change in RGB space Δx_t, and predicts the change in representation space Δz_t. Since the same pixel-level change can carry different semantic meanings depending on visual context, we condition the motion encoder on the current frame's embedding z_t via cross-attention, grounding the motion prediction in the semantic state of the current frame:

$$\Delta z_t = m_\phi(\Delta x_t;\, z_t) \tag{2}$$

Additive latent composition. With the frame encoder learning to encode the current frame in representation space and the motion encoder learning the change in representation space, predicting the next frame's representation reduces to a simple additive composition:

$$\hat{z}_{t+1} = z_t + \Delta z_t \tag{3}$$

This decomposition of $\hat{z}_{t+1}$ into z_t and Δz_t cleanly separates the goal into two objectives: the frame encoder is responsible for learning the content in a frame, and the motion encoder learns how that content evolves over time.

Preventing collapse. Supervising the next frame's predicted representation $\hat{z}_{t+1}$ requires a target: z_{t+1}, the next frame x_{t+1} encoded by the frame encoder. However, this makes the TDV recipe prone to collapse, as z_{t+1} is also produced by the same frame encoder that is being trained. Therefore, the encoder can trivially achieve near-zero loss by collapsing all representations to a constant, making the target easy to predict, but meaningless [51]. To prevent this, we adopt a teacher-student framework following DINO [17]: we maintain two copies of the frame encoder, a *student* updated by gradient descent, and a *teacher*, whose parameters are a slowly-evolving exponential moving average (EMA) of the student. The target is then produced by the teacher, which we denote as z_{t+1}^{teacher}. Both the student and teacher pass their representations through respective projection heads, and we apply a cross-entropy loss between the resulting prototype distributions. This penalizes collapse directly: if all frames map to the same representation, the distributions become identical across frames and the cross-entropy loss increases, forcing the encoder to maintain discriminative representations. The teacher's parameters evolve slowly via EMA, ensuring the student and teacher remain sufficiently different at any point in training to provide stable, non-trivial, prediction targets that the student cannot trivially satisfy by collapsing to the same distribution [16, 17]. We illustrate the complete architecture in Figure 2.

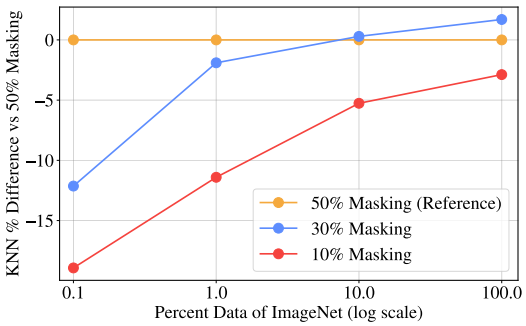

Figure 3: **Need for Assumptions Decreases as Data Scale Increases.** KNN accuracy on ImageNet-1k for three masking ratios (our proxy for assumption strength) across data scales, reported as percentage-point difference vs. 50% masking. At 0.1% data, 30% masking trails 50% masking by over 12 percentage points; by 100% data, it surpasses 50% masking. Lighter masking (10%) follows the same trend but lags, suggesting it may eventually surpass other masking ratios with increased scale. These results demonstrate that the optimal amount of inductive bias decreases with scale, motivating TDV.

3.3 TDV Training Objective

With the architecture set, we now describe how each component is supervised. TDV is trained with a weighted combination of two losses, each targeting a distinct objective.

Temporal prediction loss ($\mathcal{L}_{\text{mse}}$). The first loss directly supervises our causal principle established in Section 3.2: the motion encoder must produce a Δz_t that, when added to z_t, accurately recovers the

Table 2: **Semantic Segmentation Performance With UperNet.** We benchmark the Semantic Segmentation performance of TDV compared to iBOT and DINO on ADE20K and Cityscapes. TDV achieves competitive performance relative to iBOT and DINO despite learning without strong inductive biases.

			Semantic Segmentation (UperNet)			
			ADE20K		Cityscapes	
Method	Arch	Pretrain	↑mIoU	↑mAcc	↑mIoU	↑mAcc
iBOT [28]	ViT-S	SSv2	10.60	14.53	39.34	45.36
DINO [17]	ViT-S	SSv2	**10.71**	**14.64**	**39.85**	**45.68**
TDV	ViT-S	SSv2	10.54	14.48	37.54	43.09
iBOT [28]	ViT-B	SSv2	9.94	**13.65**	38.94	**44.31**
DINO [17]	ViT-B	SSv2	**10.48**	11.14	**39.97**	43.09
TDV	ViT-B	SSv2	9.57	10.70	36.21	42.59

next frame's embedding. This is enforced via a mean-squared error between the predicted next-frame embedding $\hat{z}_{t+1} = z_t + \Delta z_t$ and the teacher-encoded target z_{t+1}^{teacher}:

$$\mathcal{L}_{\text{mse}} = \left\| \hat{z}_{t+1} - \text{sg}(z_{t+1}^{\text{teacher}}) \right\|_2^2, \tag{4}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, ensuring this loss only updates the motion encoder and student frame encoder, not the teacher.

Self-distillation loss ($\mathcal{L}_{\text{dino}}$). The second loss addresses the collapse problem described in Section 3.2: without an additional signal, the frame encoder can trivially satisfy $\mathcal{L}_{\text{mse}}$ by collapsing all representations to the same embedding. We therefore apply a cross-entropy objective inspired by DINO [17] between student and teacher projection distributions, with one extension: we apply this loss over *both* the `[CLS]` token and *the patch tokens*, encouraging spatially consistent representations at the patch level beyond what the original DINO formulation provides. Let p_s and p_t denote the student and teacher projection distributions, normalized with temperatures τ_s and τ_t respectively (in practice, we set $\tau_t = \tau_s = 0.1$). The loss is then:

$$\mathcal{L}_{\text{dino}} = -\sum_k p_t^{(k)} \log p_s^{(k)}, \tag{5}$$

where k indexes over the K prototype dimensions of the projection head. The teacher distribution is additionally centered with a running mean to prevent dimensional collapse in the absence of temperature asymmetry [17].

Putting these together, we get the complete training objective for TDV:

$$\mathcal{L} = \lambda_{\text{mse}} \mathcal{L}_{\text{mse}} + \lambda_{\text{dino}} \mathcal{L}_{\text{dino}}, \tag{6}$$

where λ_{mse} and λ_{dino} are tunable hyperparameters.

4 Experimentation

4.1 Motivating Weaker Assumptions

To provide empirical weight to our philosophical argument—that weaker assumptions yield superior asymptotic performance as data scales—we evaluate various models across different subsets of ImageNet-1k [53]. By identifying the top-performing inductive biases at each data scale, we can observe how the optimal "strength" of assumptions shifts. Specifically, we conduct these evaluations using data subsets of 0.1%, 1%, 10%, and 100%. To measure the strength of these assumptions, we utilize masking with values of 10%, 30%, and 50% as a continuous proxy (note that these are values not for TDV, but for testing our argument regarding weaker assumptions; more details are in Section D). We use masking for two primary reasons: it allows for a granular axis, unlike discrete changes such as switching from contrastive learning to self-distillation, and it represents a clear spectrum of assumptions. For instance, requiring models to treat images with 50% masking as "similar" to their original imposes a strong assumption that only the high-level semantics remaining

Table 3: **Optical Flow and Stereo Depth Evaluation.** TDV outperforms iBOT and DINO on most optical flow and stereo depth comparisons, with a small trade-off on stereo depth average error.

			Optical Flow MPI-Sintel		**Stereo Depth** SceneFlow (final)		
Method	Arch	Pretrain	↓EPE (clean)	↓EPE (final)	↓Avg Err.	↓bad@0.5px	↓bad@1px
iBOT [28]	ViT-S	SSv2	11.31	11.27	**3.50**	65.51	44.91
DINO [17]	ViT-S	SSv2	13.03	12.92	3.64	63.25	45.30
TDV	ViT-S	SSv2	**9.84**	**10.75**	4.25	**56.89**	**39.70**
iBOT [28]	ViT-B	SSv2	11.66	11.82	**3.75**	62.49	44.18
DINO [17]	ViT-B	SSv2	11.63	**11.28**	3.91	62.97	44.64
TDV	ViT-B	SSv2	**10.97**	11.85	3.98	**54.62**	**37.33**

are sufficient for representation. Alternatively, requiring image similarity at 10% masking is a relatively weak assumption, as most of the image details remain intact.

The results for these experiments are shown in Figure 3, where the results demonstrate a strong trend between which approaches perform best and the amount of data being leveraged. With just 0.1% of ImageNet, the best performing masking ratio is 50%, with 30% and 10% masking falling behind by a significant margin. However, as the amount of data increases, 30% masking eventually outperforms 50% masking, with 10% masking approaching the performance of 50% masking. These results demonstrate that as data increases, the optimal amount of assumptions made, represented here as masking ratio, decreases. This further reinforces our motivation for TDV—to learn representations without any strong inductive biases.

4.2 Downstream Evaluations

Following [54], we argue that semantic benchmarks such as linear probing, k-NN retrieval, and action recognition probe *ventral stream* [55, 56] skills (what), and generally do not accurately measure spatial or temporal representation quality. However, understanding structure and motion, which is performed in the human brain by the *dorsal stream*, is fundamental to real-world vision applications such as robotics, autonomous driving, and 3D scene understanding. These tasks are often bottlenecked less by semantic representations and more by low-level spatial-temporal information. We therefore focus our evaluations on such properties, specifically segmentation, optical flow and stereo depth, as they demand representations to retain spatial structure and temporal correspondence, precisely the properties suppressed by strong semantic priors but preserved by TDV.

For our experimental setup, we pretrain all models on the SomethingSomethingV2 (SSV2) [57] dataset, as it has well-defined motion data and is a standard video benchmark [58]. We then evaluate all models on the downstream tasks using the pretrained backbones, with individual setup details provided in Appendix D.3.

On semantic segmentation, TDV achieves results comparable to DINO and iBOT, trailing behind by a small margin on both mIoU (mean intersection over union) and mAcc (mean per-class accuracy) as shown in Table 2. The competitiveness, visualized in Figure 4, suggests that TDV is capable of learning spatially coherent features that a segmentation head can leverage even without an explicit semantic objective. This competitiveness holds despite TDV producing less object-focused `[CLS]`-token attention than DINO and iBOT (Figure A.1), likely because segmentation relies on patch-level rather than `[CLS]`-token features. We believe the remaining performance gap likely reflects the absence of augmentations like local cropping, which may provide better semantic context.

We evaluate TDV on temporal tasks against DINO and iBOT in Table 3. On optical flow, TDV consistently outperforms both DINO and iBOT on EPE (endpoint error, the average pixel-level distance between predicted and ground truth flow vectors). We believe this can be attributed to TDV explicitly learning to predict how representations evolve between frames, which naturally preserves the local motion structure that methods trained on images with invariance augmentation tend to discard (optical flow predictions are visualized in Figure 5). On stereo depth, TDV achieves lower "bad" pixel rates at both the 0.5px and 1px thresholds across both architectures, indicating that TDV makes significantly fewer large correspondence errors than DINO and iBOT. The slightly higher average disparity error suggests that while TDV makes fewer large mistakes, it can still struggle to

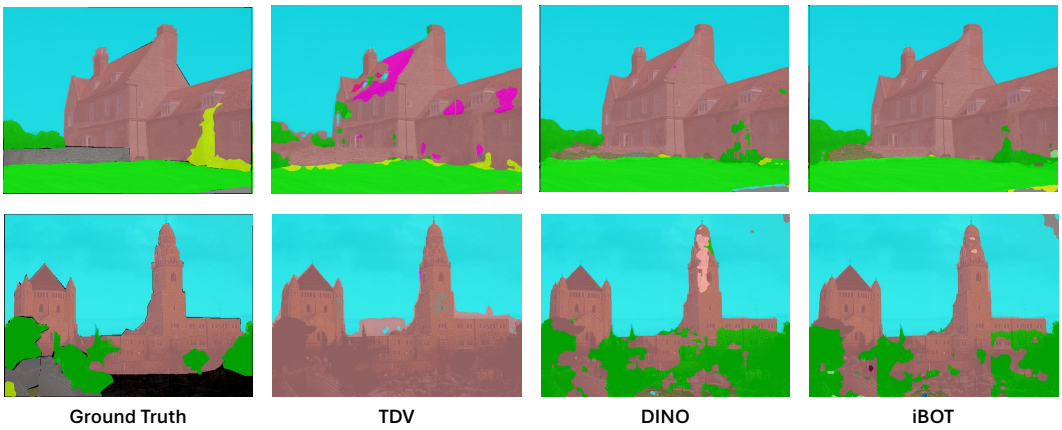

Figure 4: **Semantic Segmentation on ADE-20K (UperNet).** TDV performs competitively to DINO and iBOT, with broader region extents but slightly worse boundary separation.

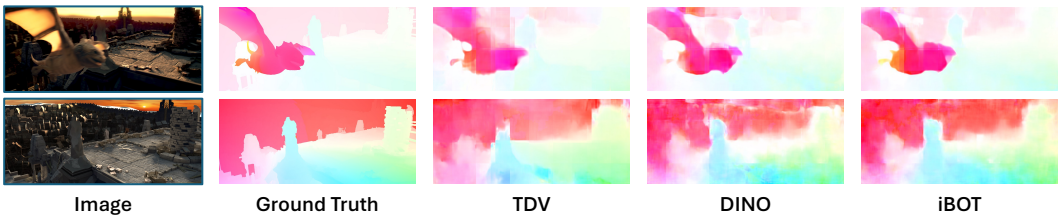

Figure 5: **Optical Flow on MPI-Sintel.** TDV produces locally consistent flow compared to DINO and iBOT, though artifacts remain in occluded regions across all methods.

recover precise depth in ambiguous regions where semantic context would otherwise help. These performance gains carry over to the features themselves: Figure B.1 shows PCA visualizations of patch-level features, where TDV produces spatially coherent feature maps.

4.3 TDV Ablation Studies

We ablate the key design choices of TDV to identify the components critical to performance and stability. We pretrain TDV on SSv2 and use online ImageNet KNN Top-5 accuracy [8] as a proxy for general performance, as it is cheap to compute during training, and gives a meaningful signal for representation quality. We show ablation results in Table 4.

The two ablations that cause training to collapse the most are removing the motion encoder and removing the MSE loss, highlighting them as two critical components of TDV. The motion encoder provides the temporal signal necessary for learning, while the MSE loss directly supervises it to predict meaningful changes in representation space. Notably, removing the motion encoder entirely and relying solely on the DINO loss across consecutive frames—effectively reducing TDV to a simple temporal invariance objective—is also not sufficient to learn representations, suggesting that explicitly modeling temporal differences between frames is a necessary choice.

Among the remaining design choices, including the `[CLS]` token in cross-attention and applying the DINO loss on the `[CLS]` token both contribute meaningfully to performance. These suggest that grounding motion predictions in a global scene representation helps the motion encoder focus on semantically meaningful changes. For the teacher's output distribution, removing centering causes a much larger performance drop than removing temperature sharpening, which we attribute to centering preventing the distribution from becoming too peaked on a single mode, a subtler form of collapse. Finally, we find that standard absolute positional encodings consistently outperform RoPE [59] across our experiments. While this ablation study addresses the components that work, we document all the other design choices and training strategies we tried that did not work in Appendix B.1.

Table 4: **Performance Impact of Key TDV Ablations.** We ablate key components of the full TDV recipe pretrained on SSv2 and report KNN (Top-5) accuracy on ImageNet-1k alongside training stability. We find that removing the motion encoder or MSE loss causes training collapse, identifying them as critical components. Similarly, centering, DINO loss on the `[CLS]` token, and cross-attention with the `[CLS]` token contribute meaningfully to performance without affecting stability.

	KNN (Top-5) ↑	Avoids Collapse
Full TDV recipe	17.05	✓
No Temperature	15.85	✓
No Centering	11.15	✓
No [CLS] in Cross Attention	10.78	✓
No Centering or Sharpening	10.68	✓
No DINO Loss on [CLS]	10.66	✓
RoPE instead of Positional Enc.	10.25	✓
No Motion Encoder	1.87	✗
No MSE Loss	1.58	✗

5 Future Works and Broader Impact

TDV's weak inductive biases and joint frame-motion encoder design open up several new directions for future work. First, deep learning approaches with weaker assumptions tend to scale more favorably with compute and data [6, 7, 10, 11]. Since TDV avoids augmentations, masking, contrastive objectives, and other strong inductive biases, it is positioned for stronger asymptotic performance than existing recipes. Second, unlike existing visual representation learning approaches, TDV relies on no vision-specific techniques, such as masking or augmentation. Therefore, we believe TDV could be applied to any modality with high temporal consistency, including audio, proprioception, and touch. Third, TDV could enable efficient video encoding. Modern approaches for representing video typically pass every frame through a full image encoder. In contrast, using TDV, only the initial frame needs the frame encoder, with subsequent frames represented by composing the previous frame's representation with a lightweight motion encoder. This is similar to classical video codecs such as MPEG, which exploit temporal redundancy by storing keyframes and inter-frame deltas.

6 Limitations and Conclusion

Limitations. While we achieved promising results with TDV in this work, there remain several limitations for further adoption. First, while TDV matches existing approaches on dense spatial tasks, it does not achieve state-of-the-art results across the board. We view this as expected for a first attempt at representation learning without strong inductive biases, and anticipate that future work can build on the recipe to close the remaining gap. Second, TDV did not achieve strong performance when measured on semantic benchmarks. We believe this is largely caused by a lack of inductive biases for learning invariances, such as local/global crops, which most existing visual representation learning approaches rely on [51]. Third, we found that scaling video data to larger video datasets than SomethingSomethingV2 [57] did not improve performance. We believe that this was caused by a lack of high-quality large scale open-source video data as well as our tuning of hyperparameters for performance on SomethingSomethingV2. We believe that with access to larger, higher-quality video datasets and better hyperparameters, TDV should scale further. Future work can search for more optimal hyperparameters that scale better to larger video datasets and model sizes.

Conclusion. In this work we proposed Temporal Difference in Vision (TDV), the first approach for learning representations from videos without any supervision, raw pixel space reconstruction, or strong inductive biases. TDV achieves comparable or sometimes improved dense/spatial task performance to state-of-the-art visual representation learning recipes such as DINO [17] and iBOT [28], while not relying on strong inductive biases. As deep learning approaches leveraging weaker assumptions generally scale better, TDV lays the groundwork for potentially more scalable representation learning.

7 Acknowledgements

We extend special thanks to Chris Hoang for his helpful discussions and advice; we are also thankful to Laude Institute for supporting this work. This work is based upon work supported by the U.S. National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 21-46756, U.S. DARPA ECOLE Program No. #HR00112390060, DARPA ITM Program No. FA8650-23-C-7316, NSF Molecule Maker Lab Institute, an AI Institute for Molecular Discovery, Synthesis Strategy, and Manufacturing funded by the U.S. National Science Foundation under Awards No. 2019897 and 2505932, the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award No. 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges, and NSF NAIRR award. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government or the National Science Foundation. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This research used the Delta and DeltaAI advanced computing and data resources, which are supported by the National Science Foundation (award OAC 2320345 and award OAC 2005572) and the State of Illinois. Delta and DeltaAI are joint efforts of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[2] 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*, pages 10684–10695, 2022. 1

[3] OpenAI. Gpt-4 technical report, 2023. 1, 5

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[5] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1

[6] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019. 1, 4, 10

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 4, 10, 22

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 5, 9, 17, 18, 22, 24

[9] 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. 2, 4

[10] Théo Moutakanni, Maxime Oquab, Marc Szafraniec, Maria Vakalopoulou, and Piotr Bojanowski. You don't need domain-specific data augmentations when scaling self-supervised learning. *Advances in Neural Information Processing Systems*, 37:116106–116125, 2024. 4, 10

[11] Hyung Won Chung. Shaping the future of AI from the history of Transformer. https://www.youtube.com/watch?v=orDKvo8h71o, 2025. Talk. 10

[12] Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers. *arXiv preprint arXiv:2507.02092*, 2025. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 4, 17

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 4, 17

[16] 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. 1, 4, 5, 6

[17] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1, 3, 4, 5, 6, 7, 8, 10, 17, 18, 22, 24

[18] Nikolaas Tinbergen. *The study of instinct*. Pygmalion Press, an imprint of Plunkett Lake Press, 2020. 2, 18

[19] Reuven Dukas. Evolutionary biology of insect learning. *Annu. Rev. Entomol.*, 53(1):145–160, 2008. 2

[20] Simon M Reader and Kevin N Laland. Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences*, 99(7):4436–4441, 2002. 2, 18

[21] Aida Gómez-Robles, Christos Nicolaou, Jeroen B Smaers, and Chet C Sherwood. The evolution of human altriciality and brain development in comparative context. *Nature Ecology & Evolution*, 8(1):133–146, 2024. 2, 18

[22] Albert Einstein et al. On the electrodynamics of moving bodies. *Annalen der physik*, 17(10):891–921, 1905. 2, 5

[23] Giacomo Mauro D'Ariano. Causality re-established. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2123), 2018. 2, 5

[24] John S Bell. On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1(3):195, 1964. 2

[25] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023. 3, 24

[26] 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. 3, 4

[27] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44, 1988. 3, 5

[28] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3, 7, 8, 10, 17, 21, 22, 24

[29] Jason Tyler Rolfe and Yann LeCun. Discriminative recurrent sparse auto-encoders. *arXiv preprint arXiv:1301.3775*, 2013. 3, 17

[30] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 3, 17

[31] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022. 4

[32] Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024. 4, 17

[33] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems*, 37:91300–91336, 2024. 4, 17

[34] 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*, pages 8748–8763. PmLR, 2021. 4, 17

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 4

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[37] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[38] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1255–1265, 2021.

[39] Adrien Bardes, Jean Ponce, and Yann LeCun. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023. 4

[40] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17131–17141, 2023. 4

[41] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1895–1904, 2021. 4

[42] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. 4

[43] Christopher Hoang and Mengye Ren. Midway network: Learning representations for recognition and motion from latent dynamics. *arXiv preprint arXiv:2510.05558*, 2025. 4, 25

[44] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. 4, 5, 17, 18

[45] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 4

[46] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks. *arXiv preprint arXiv:2111.11398*, 2021. 4, 17

[47] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 4, 17

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 5

[49] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[50] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 5

[51] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. 5, 6, 10

[52] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003. 5

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 7, 24

[54] João Carreira, Dilara Gokay, Michael King, Chuhan Zhang, Ignacio Rocco, Aravindh Mahendran, Thomas Albert Keck, Joseph Heyward, Skanda Koppula, Etienne Pot, et al. Scaling 4d representations. *arXiv preprint arXiv:2412.15212*, 2024. 8

[55] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991. 8

[56] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992. 8, 18

[57] 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 *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 8, 10, 22, 24

[58] 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. 8

[59] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 9

[60] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 17, 22

[61] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 17

[62] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 17

[63] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 17

[64] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 17

[65] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *arXiv preprint arXiv:2502.08769*, 2025. 17

[66] Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. 17, 21

[67] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 18

[68] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 18

[69] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999. 18

[70] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

[71] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012. 18

[72] Leslie G. Ungerleider. Two cortical visual systems. 1982. URL https://api.semanticscholar.org/CorpusID:142774685. 18

[73] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nature neuroscience*, 22(6):984–991, 2019. 18

[74] Olivier J Hénaff, Yoon Bai, Julie A Charlton, Ian Nauhaus, Eero P Simoncelli, and Robbe LT Goris. Primary visual cortex straightens natural video trajectories. *Nature communications*, 12(1):5982, 2021. 18

[75] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 20

[76] 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*, pages 18995–19012, 2022. 20

[77] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024. 20

[78] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 21

[79] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 21

[80] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 21

[81] 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, 2023. 24

[82] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

[83] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

[84] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. 24

[85] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 24

[86] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 24

[87] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 24

[88] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 24

[89] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 25

[90] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015. 25

[91] 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*, pages 4040–4048, 2016. 25

[92] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. 25

A Additional Intuition

To our knowledge, TDV is the first purely unsupervised visual representation learning approach that avoids *all* of the following inductive biases simultaneously: raw data reconstruction [60], aligned data with other modalities [34], hand-crafted pretext tasks [61–63], contrastive learning [14], clustering objectives [64, 65], augmentations [14, 17], explicit invariances [44], redundancy reduction (i.e., Variance or Covariance regularization [44]), and cropping or masking [8, 66]. While each of these techniques has driven significant progress in representation learning, each also introduces assumptions that can become limiting as data and compute scale. We discuss the limitations of each below; prior work has raised similar concerns regarding how invariances and pretext biases can affect performance [46, 47].

Image Augmentations. Augmentation-based approaches [8, 14, 17] often pull together differently augmented views of the same image, implicitly treating whatever the augmentation changed as irrelevant to semantics. However, which factors are irrelevant depends on the downstream task. Color jitter encourages invariance to color, which can help coarse classification but is unhelpful for tasks where color is informative, such as bird species, ripeness, or material recognition. Similarly, spatial augmentations encourage invariance to location, which is precisely the information that detection must preserve. Any forced invariance therefore tends to help some tasks while hurting others, and we expect this trade-off to become more pronounced at the limit of learning [46, 47].

Masking. Masking-based approaches [8, 28, 66] face a closely related issue: they encourage an image and a heavily masked version of it to map to nearly the same representation, even though much of the visual content has been removed. This can collapse semantically distinct inputs together and reduce spatial awareness. Detection illustrates the concern: two images of the same object at different positions should ideally produce different representations so that location can be recovered, whereas masking pushes them toward a shared representation. Section 4.1 provides an empirical version of this argument, showing that the optimal masking ratio decreases as data scale grows.

Contrastive Learning. Contrastive approaches [14, 15] push randomly sampled images apart in representation space. This is only approximately valid, as two random images from a natural distribution often share content (e.g., two outdoor scenes, two dogs, or two faces), and separating them can discard useful structure. Contrastive learning also relies on large numbers of negative samples, which makes scaling costly in both batch size and memory.

Raw Pixel Reconstruction. Raw reconstruction-based approaches [29, 30, 60] require the representation to retain enough information to rebuild every pixel, including texture, lighting, and background detail that may be unrelated to the scene content. Both theory [32] and experiments [33] indicate that pixel-space targets yield features that are less useful for perception than latent-space targets. Requiring the representation to encode all such detail can also limit how abstract it becomes, since fine-grained appearance must be retained somewhere.

Cross-Modal Alignment. Approaches such as CLIP [34] learn vision representations by aligning them with text. This works well when paired data is plentiful, but it ties the visual representation to whatever the accompanying text describes. Captions tend to emphasize objects and actions while omitting texture, geometry, lighting, and spatial layout, so the resulting features can be biased toward nameable content and weaker on dense spatial properties that text rarely describes. The approach also requires paired modalities, which is itself a strong data assumption.

Hand-Crafted Pretext Tasks. Pretext objectives such as jigsaw puzzles [61], rotation prediction [62], and colorization [63] solve a synthetic problem in the hope that the features learned along the way transfer. A limitation is that the model only needs to learn whatever is sufficient for that specific task: rotation prediction, for instance, can rely on a few orientation cues such as sky position or text, leaving the rest of the representation underdeveloped. More broadly, the designer must anticipate what makes a useful pretext, and each choice reflects a particular view of which visual structure matters.

Clustering Objectives. Clustering-based approaches [64, 65] assign images to discrete prototypes and train the encoder to be consistent with those assignments. This assumes that the data falls into

a fixed number of well-separated clusters, which is only approximately true for natural images: semantic categories overlap, vary continuously, and exist at multiple granularities simultaneously. The number of clusters becomes a hyperparameter that bounds the granularity the representation can reach, and clustering pipelines often require additional machinery (e.g., Sinkhorn balancing or queue tricks) to avoid degenerate solutions.

Explicit Invariances and Redundancy Reduction. Methods such as VICReg [44] avoid contrastive negatives by directly regularizing feature statistics across a batch: an invariance term between augmented views, together with variance and covariance terms computed over batch dimensions to prevent collapse. The invariance term shares the limitation of augmentation-based methods, as features are encouraged to ignore whatever the augmentation changed. The variance and covariance terms introduce a separate consideration: they assume that a batch of independently sampled images should produce features that are well-spread and decorrelated, which holds only when batches are large and diverse enough to approximate the data distribution. Small batches, batches with correlated content (e.g., consecutive video frames or images from the same scene), or any setting where the i.i.d. batch assumption does not hold can push these regularizers toward the wrong target. This couples the learning signal to batch composition in a way that the underlying representation problem need not depend on.

Cropping. Random cropping [8, 17] is often treated as a default, but it constitutes a strong inductive bias. It assumes that two crops of the same image should produce similar representations, which in turn assumes both crops contain the same semantic content. On object-centric datasets such as ImageNet this generally holds; on cluttered or scene-level images, however, two crops may capture entirely different objects, and the model is then trained to treat them as equivalent. This can bias representations toward whatever survives random cropping—typically the dominant object—and away from full-scene or denser understanding.

A.1 Broader Philosophical Intuitions

We argue that the ideal inductive bias is the weakest one that still enables learning within a single lifetime of experience—strong enough to bootstrap learning from finite data, but weak enough to not bottleneck what can ultimately be learned. This mirrors biological evolution, where more capable species hardcode less behavior into their genome and instead learn from experience [18, 20, 21]: evolution has converged on the weakest priors that still permit survival-level learning within a lifetime.

Under this minimal-prior view, TDV can be understood as simply performing compression—the motion encoder captures only the change between frames, and the frame encoder captures only what is needed to predict the next frame given that change. No assumptions about augmentation invariances, negative pairs, or pixel-level fidelity are imposed. The model is simply compressing temporal experience into representations sufficient to predict the future. We view this as the minimal assumption necessary for representation learning.

Another perspective on TDV is that it is learning a (small) latent action world model [67, 68], where the motion encoder captures temporal differences (latent actions), that cause the future frame to be predictable.

A.2 Neuroscientific Intuitions

It happens that TDV maps reasonably well onto several established theories of biological vision. The MSE objective—predicting the next-frame embedding from the current one—can be viewed as an instance of *predictive coding*, where the cortex learns by minimizing errors between predicted and observed sensory input [69–71]. The factorization into a frame encoder and a motion encoder operating on Δx also loosely mirrors the dorsal/ventral dissociation in primate visual cortex [56, 72], where the dorsal stream is fed by the magnocellular pathway, itself selective for motion and temporal change. Most directly, our additive composition $\hat{z}_{t+1} = z_t + \Delta z_t$ echoes the *temporal straightening* hypothesis [73, 74], which posits that the visual system maps video onto straighter latent trajectories so that future states can be predicted by near-linear extrapolation—precisely the structure TDV imposes. While these connections did not directly motivate TDV, we find it encouraging that its design choices align with mechanisms already studied in biological vision.

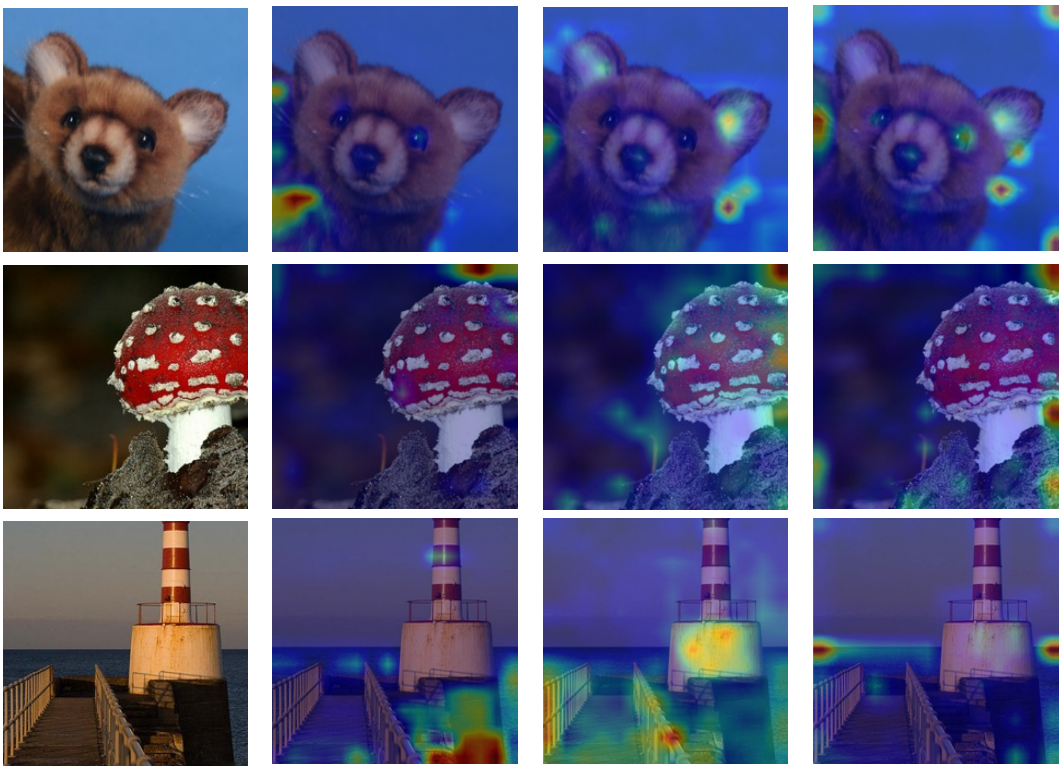

Figure A.1: **Attention Maps Reflect Pre-Training Objectives.** We visualize self-attention maps for the `[CLS]` token of ViT-B models pre-trained on SSv2 for three ImageNet images (left). Warmer colors indicate higher attention. TDV does not use a dominant `[CLS]`-token objective during pre-training, hence its `[CLS]` attention is less object-focused than that of DINO and iBOT.

Table B.1: **TDV Lags Behind DINO and iBOT on Semantic Evaluations.** We evaluate ImageNet-1k KNN Top-1 accuracy and SSV2 action-recognition Top-1 accuracy for DINO and iBOT with all augmentations enabled compared to TDV (all trained on the Something-Something V2 video dataset). TDV, being trained without any strong inductive biases, does not achieve the same semantic performance as iBOT or DINO.

		ImageNet Classification		**SSv2 Action Recognition**
Method	Architecture	KNN (Top-5) ↑	Linear (Top-5) ↑	Linear (Top-5) ↑
iBOT	ViT-S	33.46	41.29	20.30
DINO	ViT-S	34.81	44.19	19.50
TDV	ViT-S	14.74	17.52	10.10
iBOT	ViT-B	38.75	46.10	21.60
DINO	ViT-B	40.89	49.38	20.50
TDV	ViT-B	17.05	16.14	10.10

B Additional Experiments

We report results for semantic benchmarking in Table B.1, demonstrating how TDV currently lags behind SOTA recipes at semantic representation. This is likely due to a lack of strong inductive biases.

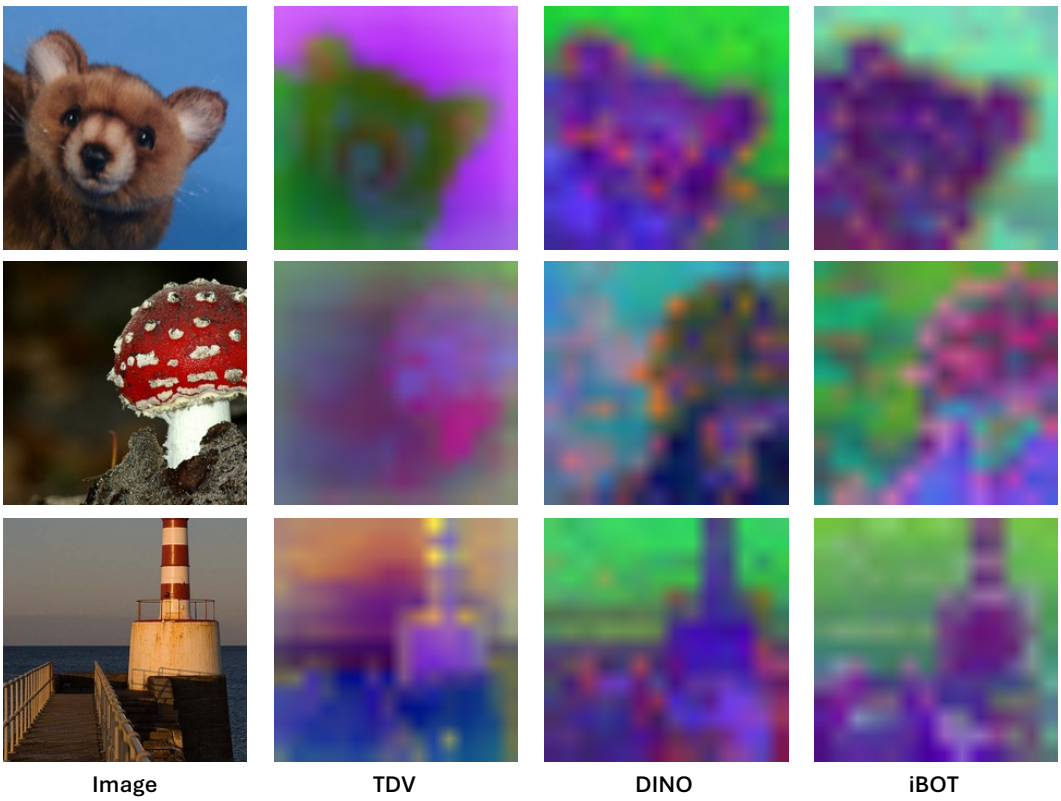

Figure B.1: **PCA Visualization Shows Patch Features Capture Coherent Object Structure.** We show RGB visualizations of the top-3 PCA components of patch-level features from ViT-B models pre-trained on SSv2, for three ImageNet images (left). Each color corresponds to a principal direction of the patch-feature space, so similar colors mark patches with similar representations. TDV produces clean, spatially coherent feature maps that align with object boundaries, often better than DINO and iBOT. This shows that TDV learns strong patch-level representations, consistent with its dense-prediction performance (Tab. 3).

B.1 Experiments We Tried that Didn't Work

Inspired by [75], we document design choices and training strategies that we explored but found to be ineffective. It's worth noting that because TDV doesn't leverage any strong inductive biases, in general, training it with current tools is not easy.

Scaling to larger video datasets. We explored pretraining on Ego4D [76] and FineVideo [77] as alternatives to SSv2, to test if more data would improve representation quality. Ego4D is an egocentric dataset with significant variance in motion where some clips are nearly static while others have fast, erratic camera motion, which we found made the RGB difference signal noisy and difficult for the motion encoder to learn from consistently. FineVideo contains many abrupt scene cuts where the difference between consecutive frames reflects an editing transition rather than natural motion. We preprocessed it into smaller chunks to avoid exposing the model to cross-scene differences, but this substantially reduced the usable data volume. Despite our preprocessed FineVideo having approximately $2\times$ more video data than SSv2, pretraining on it for the same 200k steps yielded a lower ImageNet KNN Top-5 accuracy (10.75% vs. 17.05% with SSv2). A good finding was that if you kept training TDV on FineVideo for longer, the representation quality keeps consistently improving, reaching equal KNN performance (16.04%) to SSv2 at around 600,000 steps. This suggests that the data quality and motion coherence also matter for the current TDV architecture. Future work can address making TDV more robust to noisy data.

Combining multiple datasets. We experimented with mixing SSv2, Kinetics-400 [78], and Ego4D in various combinations to increase training data diversity. In all cases, representation quality as measured by ImageNet KNN degraded relative to training on individual datasets alone. We attribute this to the heterogeneous motion statistics across datasets: Kinetics clips tend to be short and action-centric, while Ego4D has highly variable motion rates, and mixing these with SSv2 appears to make the motion prediction task less coherent rather than more informative. Future work can search for hyperparameters that generalize better to multiple datasets.

Alternative conditioning mechanisms for the motion encoder. The motion encoder in TDV is conditioned on the frame encoder's `[CLS]` + patch tokens via standard cross-attention. We explored replacing this with feature-wise linear modulation (FiLM) [79], AdaLN, AdaLN-Zero [80], and Gated AdaLN as conditioning mechanisms. All four alternatives showed promising KNN accuracy during the first few training epochs—in all cases doing 2x better than standard cross attention—but then plateaued or collapsed after a few epochs of training. The best observed KNN Top-5 values were FiLM: 7.92%, AdaLN: 8.97%, AdaLN-Zero: 10.21%, and Gated AdaLN: 9.91% vs Cross Attention: 5.4% (at that epoch). We hypothesize that while these modulation-based approaches provide stronger conditioning signal, they make it easier for the model to find degenerate solutions, directly affecting the training stability. Future work can leverage newer anti-collapse solutions, e.g. SIGReg [66].

RGB difference thresholding. We experimented with skipping the motion encoder pass for frames where the magnitude of the RGB difference fell below a threshold, with the intention of filtering out near-static frames where the motion signal is near-zero. In practice this did not improve results. We suspect that static frames still contribute useful learning signal by requiring the motion encoder to produce a near-zero delta, which may provide a natural calibration for the prediction objective.

Motion embedding divergence loss. We explored adding an auxiliary loss to push consecutive frame encoder embeddings further apart, with the intuition that a larger gap in representation space would give the motion encoder more room to learn a meaningful delta. In practice, this caused the variance of frame representations to explode and destabilized training without improving KNN or downstream performance.

iBOT-style grid masking. We attempted to combine TDV with iBOT-style [28] patch-level masked prediction, to see if adding more augmentations improves spatial performance. The main challenge is in handling masking consistently between the current frame and the RGB difference input: simply masking the same patch positions from both inputs causes collapse for masking rates above 10%. Below that threshold, performance was not meaningfully improved over the unmasked baseline, suggesting the interaction between masking and the motion objective requires more careful design than a simple direct combination.

DINO-style augmentations. We experimented with applying standard DINO augmentations (random resized cropping, color jitter, Gaussian blur, solarization) to both the current and next frame before feeding them to the encoders. This consistently led to collapse. Our interpretation is that applying strong augmentations to consecutive frames changes the relationship between them in ways that are inconsistent with the causal motion objective: if two frames are independently cropped to different spatial regions, the RGB difference no longer reflects actual scene motion, removing the core learning signal. More careful augmentation strategies that preserve the temporal relationship between frames may be possible with additional tuning.

Asymmetric teacher sharpening. DINO uses a significantly lower softmax temperature for the teacher than the student, effectively sharpening the teacher's output distribution relative to the student's. We explored applying the same asymmetry to TDV's EMA teacher. This consistently performed worse than using the same temperature for both; the best results were obtained when teacher and student were sharpened identically. We do not have a strong explanation for this difference, but note that TDV's teacher plays a different role than DINO's—it supervises next-frame predictions rather than augmented-view agreement—which may be the reason for the change in dynamics of temperature asymmetry.

Per-epoch teacher reinitialization. We experimented with reinitializing the teacher frame encoder from the student's latest checkpoint at the start of each epoch, rather than maintaining a continuous

Table C.1: **TDV Requires No Hand-Crafted Augmentations.** DINO and iBOT both rely on heavy augmentations. TDV, by contrast, uses none of these inductive biases, while still avoiding collapse. Rather, TDV learns from the natural change between frames in a video.

Inductive Bias / Augmentation	DINO	iBOT	TDV
Multi-crop (global + local)	✓	✓	×
Random resized crop	✓	✓	×
Random horizontal flip	✓	✓	×
Color jitter	✓	✓	×
Gaussian blur	✓	✓	×
Solarization	✓	✓	×
Masked image modeling	×	✓	×
Temporal frame sampling	×	×	✓

EMA. This was motivated by a concern that a slowly-drifting EMA teacher might provide stale targets as training progresses. In practice, this performed worse than a fixed high EMA momentum, likely because abrupt teacher resets remove the smoothing that EMA provides and introduce instability in the supervision signal.

Using full frames instead of RGB differences. We tested feeding the full current frame (rather than the RGB frame difference) to the motion encoder, so that both encoders receive complete image inputs. In this setting, the motion encoder has no strong inductive bias toward encoding temporal change, and in practice the model collapsed—the motion encoder tended to replicate the frame encoder's representations rather than learn complementary motion information.

Smaller motion encoders. We ablated the size of the motion encoder across a range of parameter counts. Smaller motion encoders consistently yielded lower ImageNet KNN accuracy, with a roughly monotonic relationship between encoder capacity and representation quality. This suggests the motion encoder needs sufficient capacity to model the full range of temporal changes present in video, and undersizing it creates a bottleneck in the learning signal reaching the frame encoder.

Continued pretraining of existing vision encoders. We explored using TDV to improve existing pretrained vision encoders (MAE [60] and DINOv2 [8]) by initializing the frame encoder from their weights and continuing pretraining with small learning rates while jointly training a motion encoder. With the frame encoder unfrozen, this consistently degraded the pretrained representations rather than improving them. However, keeping the pretrained frame encoder *frozen* and training only the motion encoder does work: the motion encoder successfully learns to predict frame embedding differences, recovering approximately 90% of the embedding delta for MAE and 60% for DINOv2. This suggests TDV's motion objective is compatible with existing pretrained features as a fixed representation backbone. This could be useful for efficient video encoding.

C TDV Details

C.1 Training Recipe

Algorithm 1 summarizes the TDV training procedure for a single step. The student frame encoder and motion encoder are updated by gradient descent; the teacher frame encoder is updated only via EMA and receives no gradients. We also summarize the differences between TDV and DINO/iBOT in Table C.1.

C.2 Pretraining Setup

All models—TDV, DINO [17], and iBOT [28]—are pretrained on the Something-Something V2 (SSv2) [57] dataset. SSv2 consists of approximately 220,000 short egocentric video clips depicting hand-object interactions, making it a practical choice for initial experimentation due to its manageable size and consistent quality of motion. All models are trained using ViT-S and ViT-B [7]

Algorithm 1 TDV Training Step

Require: Student frame encoder f_θ, motion encoder m_ϕ, teacher frame encoder $\bar{f}$ (EMA of f_θ)

Sample consecutive frames $x_t,\ x_{t+1}$ from a video

$z_t \leftarrow f_\theta(x_t)$ ▷ encode current frame

$\Delta x_t \leftarrow x_{t+1} - x_t$ ▷ RGB difference

$\Delta z_t \leftarrow m_\phi(\Delta x_t \mid z_t)$ ▷ encode motion, conditioned on z_t

$\hat{z}_{t+1} \leftarrow z_t + \Delta z_t$ ▷ predict next frame representation

$\bar{z}_{t+1} \leftarrow \bar{f}(x_{t+1})$ ▷ teacher target (no gradient)

$\mathcal{L} \leftarrow \lambda_{\text{mse}} \underbrace{\|\hat{z}_{t+1} - \bar{z}_{t+1}\|_2^2}_{\text{MSE over all tokens}} + \lambda_{\text{dino}} \underbrace{\text{DINOLoss}(\hat{z}_{t+1},\ \bar{z}_{t+1})}_{\text{cross-entropy on all tokens}}$

Update $\theta,\ \phi$ via backpropagation

$\bar{\theta} \leftarrow \tau\,\bar{\theta} + (1-\tau)\,\theta$ ▷ EMA update teacher

Table C.2: **Pretraining Hyperparameters.** Hyperparameters used for TDV, DINO, and iBOT pretraining on SSv2.

Hyperparameter	TDV	DINO	iBOT
Architecture	ViT-S/B	ViT-S/B	ViT-S/B
Patch size	14	16	16
Epochs	20	20	20
Batch size (Images)	256	256	256
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-4	5e-4	5e-4
LR schedule	cosine	cosine	cosine
Warmup epochs	0.5	10	10
Weight decay	0.01	0.04	0.04
EMA momentum (τ)	0.99	0.996	0.996
Student temperature (τ_s)	0.1	0.04	0.04
Teacher temperature (τ_t)	0.1	0.1	0.1
Projection head dim	32768	1024	8192
λ_{mse}	1.5	—	—
λ_{dino}	1.5	—	—

architectures. Each model is trained for around 200,000 steps (20 epochs), and we report results from the final checkpoint. DINO and iBOT are pretrained on SSv2 using their respective objectives and augmentations tuned, which enables their maximum performance on SSv2.

C.3 Hyperparameters

Model and optimization. Table C.2 lists the hyperparameters used to train all three models. Shared settings—batch size, optimizer, learning rate schedule, and EMA momentum—are kept identical across methods wherever possible to ensure a fair comparison. Other hyperparameters related to customized augmentations and temperature sharpening were kept at their default values to get the best performance from all recipes.

Data and temporal sampling. Table C.3 lists the data and temporal sampling hyperparameters specific to TDV. The key input is the RGB difference $\Delta x_t = x_{t+1} - x_t$, computed between two temporally adjacent frames sampled at a fixed stride. The stride controls how much motion is visible in the differences between images: too small a stride produces near-zero differences for slow-moving scenes, while too large a stride introduces large, incoherent pixel jumps where object positions change so drastically that the difference no longer captures meaningful motion structure.

Table C.3: **Data Hyperparameters.** Hyperparameters for Data and Temporal Sampling for TDV pretraining

Hyperparameter	Value
Dataset	SSv2
Input resolution	224×224
Frames sampled per clip	16
Time Between Frames	0.25
RGB difference clipping	no
Spatial cropping	center crop only
Horizontal flip	no
Color jitter / augmentations	none
Masking	none

D Experimental Details

D.1 Philosophical Backing Experiments

To empirically support our argument that weaker assumptions perform better as data scales, we train a series of models on subsets of ImageNet-1k [53] and measure how performance varies with both data scale and augmentation strength. We use the DINO [17] codebase and augment it with iBOT-style [28] patch-level masked prediction, varying the masking ratio as a continuous proxy for assumption strength. To prevent collapse under low-augmentation regimes, we retain random resized cropping with two global crops (consistent with DINO's default setup); all other augmentations are disabled. We vary the grid masking ratio across $\{10\%, 30\%, 50\%\}$ and train each of these masking configurations on data subsets of $\{0.1\%, 1\%, 10\%, 100\%\}$ of ImageNet-1k. We evaluate each model using ImageNet KNN Top-1 accuracy and report the results in Figure 3.

D.2 Semantic Evaluations

In addition to the more semantic/temporally focused representation benchmarking conducted in Section 4, we conduct experiments aimed at measuring the semantic representations. Because the default TDV training recipe does not leverage augmentations, explicit invariances, or any strong inductive biases, it is expected for learned representations to not be very semantic [81–84]. Hence, the expected performance on semantic tasks without these inductive biases is not great. We confirm this in Table B.1, where the semantic performance of the default TDV recipe lags behind existing models.

Action recognition. We evaluate action recognition on Something-Something V2 [57] following the frozen evaluation protocol of V-JEPA [25]: we freeze the pretrained encoder and train a task-specific linear probe on top of the `[CLS]` token representations extracted from 8 uniformly sampled frames per video. Frame features are concatenated along the temporal dimension before being passed to the probe head. We report the Top-5 accuracy for action recognition on the validation set for SSv2.

KNN on ImageNet. To monitor representation quality and detect collapse during pretraining without incurring the cost of full downstream evaluation, we compute an online ImageNet KNN Top-5 accuracy [8] after each training epoch. At each evaluation step, we extract `[CLS]` token features for all of ImageNet training subset images using the current student encoder and teacher encoders, and perform k-nearest-neighbor classification ($k = 20$) in feature space. This provides a lightweight signal that correlates well with final representation quality and allows us to detect collapse early, as collapsed representations yield worse than near-chance KNN accuracy. We report accuracy for ViT small and base variants of iBOT, DINO and TDV in Table B.1

D.3 Downstream Spatial Evaluations

Semantic segmentation. We evaluate semantic segmentation using UperNet [85] via the MMSegmentation [86] toolbox on ADE20K [87] and Cityscapes [88] under a frozen backbone evaluation protocol: the pretrained encoder weights are frozen and only the UperNet segmentation head is

trained. We use the standard UperNet configuration for ViT backbones and train the segmentation head for 320,000 steps for all models. We report mIoU and mAcc on the respective validation sets using the checkpoint from the final training step. All baselines use the same configuration.

Optical flow. We follow the evaluation protocol of CroCo [89] and perform full end-to-end fine-tuning on FlyingChairs [90], FlyingThings3D [91] and MPI-Sintel [92] training sets, then evaluate on MPI-Sintel [92] clean and final validation sets using endpoint error (EPE, lower is better). For the decoder, we use the two-frame feature fusion module from Midway Networks [43], which processes features from two consecutive frames before passing them to a DPT prediction head (standard CroCo setup). This decoder architecture is identical for all three methods (TDV, DINO, iBOT) to ensure fair comparison. We fine-tune all models for 240 epochs using the default CroCo hyperparameters and report results from the final checkpoint.

Because the TDV architecture naturally contains a motion encoder with features that may contain richer temporal correspondences than frame encoder features alone, we additionally experiment with supplying intermediate representations from TDV's motion encoder as an additional input to the DPT head. Removing the Midway Networks[43] decoder and passing embeddings from intermediate layers of the motion encoder directly to the DPT head result in EPE (clean) 14.53 and EPE (final) 14.52 for the TDV base variant and EPE (clean) 14.79 and EPE (final) 14.39 for the TDV small variant. While these results are currently worse than using the Midway Networks decoder, careful selection of intermediate layers from frame and motion encoder as input to the DPT and more hyperparameter tuning can improve performance directly. We leave this analysis for future work.

Stereo depth. We follow the same CroCo [89] evaluation protocol for stereo depth: full end-to-end fine-tuning on the SceneFlow (final) [91] training set, evaluated on the SceneFlow final validation set. We use the same Midway Networks [43] two-frame decoder and DPT head as in the optical flow evaluation. We fine-tune all models for 32 epochs using default CroCo hyperparameters and report average disparity error and bad pixel rates at 0.5px and 1px thresholds from the final checkpoint.

As with optical flow, we experiment with using intermediate representations from TDV's motion encoder as additional input to the DPT head. This variant results in average disparity error 6.93 for the TDV base variant and 7.23 for the TDV small variant.

D.4 Compute Resources

Pretraining. All TDV, DINO, and iBOT pretraining runs were conducted on 2 NVIDIA H100 GPUs (80 GB GPU memory each). Each pretraining run trains for 20 epochs on SSv2 and takes approximately 48 hours.

Semantic segmentation fine-tuning. UperNet fine-tuning on ADE20K and Cityscapes was run on a single NVIDIA H100 GPU (80 GB GPU memory) for 320,000 steps, taking approximately 20 hours per run.

Optical flow fine-tuning. Optical flow fine-tuning on FlyingChairs, FlyingThings3D, and MPI-Sintel was run on 2 NVIDIA H100 GPUs (80 GB GPU memory each) with BF16 precision enabled in the CroCo codebase, taking approximately 48 hours per run.

Stereo depth fine-tuning. Stereo depth fine-tuning on SceneFlow was run on the same 2 NVIDIA H100 GPUs (80 GB GPU memory each) for 32 epochs, taking approximately 16 hours per run.