LeVJEPA: Efficient & Scalable Video Pretraining without the Heuristics

Anonymous authors
Paper under double-blind review

Abstract

Video carries the temporal structure of the physical world, yet learning representations from it remains computationally expensive. Current self-supervised methods either prevent representation collapse through architectural asymmetries (an exponential-moving-average target encoder, a stop-gradient, and a predictor), or avoid it by reconstructing masked content in pixel space with a dedicated decoder. We introduce **LeVJEPA**, the first extension of LeJEPA to video. A single encoder is trained with an invariance loss between global and local views of a clip, while collapse is prevented by SIGReg with provable guarantees. Thus, neither architectural asymmetries nor pixel reconstruction is needed. This formulation has two properties that we study in this paper. First, the pretraining cost depends only on the number of tokens the encoder observes. Dropping tokens uniformly at random reduces this number and, at the same time, improves downstream accuracy. At matched epochs on identical data, LeVJEPA matches or outperforms V-JEPA 2 across ViT-S/B/L with 5.6 to $20.8\times$ less pretraining compute. At matched FLOPs, it exceeds the strongest video baseline by 7.6 points on ImageNet-1K while remaining competitive on motion-centric benchmarks. Second, since no target encoder or predictor is involved, the encoder can be trained with block-causal attention without loss in accuracy, so that the representation of each frame depends only on past frames. Finally, compared to DINOv2 trained on frames of the same videos with matched compute, LeVJEPA remains within 3.1 points on ImageNet-1K while nearly doubling the accuracy of DINOv2 on Something-Something-v2. These results suggest that, once its computational overhead is removed, video is a viable substrate for general-purpose visual pretraining.

1 Introduction

Video is a natural source for learning representations of the physical world. It is abundant, requires no annotation, and carries temporal structure (motion, causality, object permanence) that static images cannot provide (Spelke et al., 1995; Rao & Ballard, 1999). Yet video pretraining has remained the expensive option for visual representation learning. A single clip contains an order of magnitude more tokens than an image, and the mechanisms that current methods use to avoid representation collapse add further cost on top: leading video joint-embedding methods maintain two encoders and a predictor (Grill et al., 2020; Bardes et al., 2024), and inherit masking schemes designed around this architecture rather than around video itself (Tong et al., 2022). In this work, we train a video encoder with SIGReg (Balestriero & LeCun, 2025), a regularizer that excludes collapse with a provable guarantee. This makes the previous collapse-prevention mechanisms, as well as the design choices built around them, unnecessary. As we show in this work, removing them considerably simplifies video pretraining and reduces its cost.

Pretraining then reduces to a single encoder trained with a single loss, comprising 1) an invariance term between a global view and several local views of a clip, and 2) SIGReg applied to their embeddings. Prior methods spend a large part of each training step on components that only serve to stabilize learning, such as a target-encoder forward pass over the full clip or a predictor over masked queries. In our case, instead, every operation contributes to the objective. As a result, the cost of a step depends only on the number of tokens the encoder sees.

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

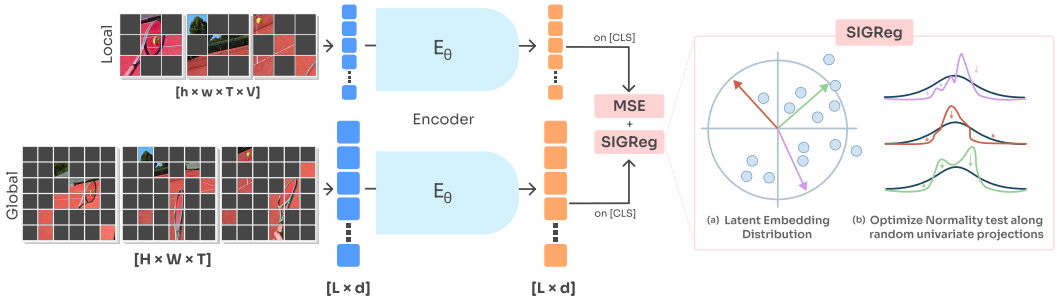

Figure 1: **LeVJEPA training.** From each clip we construct one global view ($H \times W \times T$) and several local views ($h \times w \times T \times V$). All views share the same temporal window, and local views are additionally cropped and photometrically augmented. Each view is tokenized into patch tokens, of which 95% are dropped uniformly at random (grey); the L retained tokens ($L \times d$) are prepended with a learnable `[cls]` token to form the encoder input. All views are processed by the same block-causal encoder E_θ, and only the `[cls]` embedding of each view enters the loss. The objective combines a mean squared error (MSE) that drives each local `[cls]` embedding toward the global one (with gradients flowing through all views), and SIGReg (Balestriero & LeCun, 2025) *(right)*. SIGReg (a) projects the embeddings onto random directions, and (b) penalizes the deviation of each projection from a standard Gaussian via a normality test. In this way, the embedding distribution is constrained to an isotropic Gaussian, which provably prevents collapse.

This has two consequences. First, the set of observed tokens becomes a free parameter and we find that it can be made very sparse. The resulting encoder reaches accuracy comparable to state-of-the-art video joint-embedding methods at a fraction of their pretraining FLOPs. Second, since neither a target encoder nor a predictor is involved, the attention pattern of the encoder is not constrained. We exploit this freedom to train with block-causal attention, bidirectional within a frame and causal across frames. Causality aligns with the predictive feature principle (Rao & Ballard, 1999), which requires frame representations to be computable from past observations only. At inference time, causality allows the representation of a video to be updated as new frames arrive, without re-encoding past frames (a key requirement of autoregressive world models and streaming settings). In contrast, bidirectional encoders need to re-encode the whole clip or rely on a separate temporal model trained after pretraining (Zhou et al., 2025; Assran et al., 2025; Maes et al., 2026).

We call this method **LeVJEPA**, the first extension of the LeJEPA paradigm to video. We evaluate it with frozen probing against video and image pretraining baselines retrained on the same data, both at matched epochs and at matched FLOPs. Our experiments support three findings:

- *FLOP-efficient video pretraining.* At matched epochs, LeVJEPA matches or outperforms V-JEPA 2 across model sizes with up to 20.8× less pretraining compute (Figure 3). At matched FLOPs, it outperforms video baselines on ImageNet-1K (IN1K) and Kinetics-400 (K400) while remaining competitive on Something-Something-v2 (SSv2) (Table 2). Compared to DINOv2 trained on frames of the same videos at equal compute, it approaches DINOv2 on IN1K while reaching almost twice its accuracy on SSv2 (Table 3). Pretraining is also feasible on a single consumer GPU (Section 5.3).
- *Causal frame representations at no accuracy cost.* Block-causal attention matches bidirectional attention under frozen probing (Table 1). Thus, temporal ordering becomes a property of the encoder itself, and frame representations can be extended to incoming frames without re-encoding and without a separate temporal model.
- *A simpler video pretraining recipe.* The architecture reduces to an encoder and a small projector, with no predictor network, target encoder, or stop-gradient, and the single hyperparameter of the objective is fixed to its published default in all experiments. Other design choices simplify as well. Uniform random token dropping acts as an augmentation, with accuracy increasing as more tokens are dropped (Figure 2), and outperforms structured tube masking (Section 4). Per-frame tokenization outperforms temporal patch aggregation at the same token budget (Table 1). Finally, semantically organized patch representations emerge although only the `[cls]` token is supervised (Figures 4 and 5).

2 Related Work

Self-supervised learning from images. Joint-embedding methods map augmented views of the same image to nearby representations, and mainly differ in how they prevent collapse to a constant solution. BYOL (Grill et al., 2020) showed that negative pairs are not needed when an online network with a predictor head regresses the output of an exponential moving average (EMA) target network with a stop-gradient. DINO (Caron et al., 2021) adopts a similar teacher–student setup with centering and sharpening, and introduces multi-crop, where several small local views are matched to a few global ones. DINOv2 (Oquab et al., 2024) combines this principle with masked-image modeling (Zhou et al., 2022) and scales it to curated data, while I-JEPA (Assran et al., 2023) replaces hand-designed augmentations with prediction between masked regions in representation space. LeV-JEPA adopts the global–local view construction of these methods, but none of their mechanisms for preventing collapse, instead replacing them with principled SIGReg regularization.

Explicit constraints on the embedding distribution. Other methods prevent collapse by constraining the embeddings directly. VICReg (Bardes et al., 2021) and Barlow Twins (Zbontar et al., 2021) combine an invariance term with variance and covariance terms, and require neither a target network nor a stop-gradient. However, they only constrain second-order statistics and do not specify which embedding distribution is desirable. LeJEPA (Balestriero & LeCun, 2025) addresses this point: under mild assumptions, the isotropic Gaussian uniquely minimizes the worst-case downstream probing risk. To enforce this distribution, SIGReg projects the embeddings onto random directions and tests whether each one-dimensional projection is Gaussian, at a cost that scales linearly with batch size and embedding dimension. The resulting approach uses a single network and a single loss, but has so far only been applied to static images.

Pixel reconstruction from video. Masked autoencoding reconstructs missing content in pixel space and, as this objective has no trivial solution, does not require any mechanism to prevent collapse. VideoMAE (Tong et al., 2022) observes that random masking makes the task too easy, since masked patches can be inferred from the same location in adjacent frames. To remove this shortcut, the authors propose tube masking, which occludes the same spatial region in every frame and allows masking ratios of 90–95%. VideoMAEv2 (Wang et al., 2023) scales this approach with an additional mask on the decoder side. Tube masking is thus designed to make reconstruction non-trivial. However, when nothing has to be reconstructed, as in our setting, we find it to be harmful (Section 4).

Feature prediction and video world models. V-JEPA (Bardes et al., 2024) predicts representations instead of pixels. An encoder and a narrow predictor are trained jointly to predict the representations of masked regions, with targets produced by an EMA copy of the encoder. As in VideoMAE, the masks span the full temporal extent of the clip. V-JEPA 2 (Assran et al., 2025) scales this approach in data, model size, schedule, and resolution, and reports state-of-the-art frozen performance on motion understanding and action anticipation. For planning, the action-conditioned variant V-JEPA 2-AC freezes the encoder, applies it per frame, and trains a separate block-causal predictor conditioned on actions and end-effector states on robot interaction data. This enables zero-shot goal-conditioned manipulation via model-predictive control. Other works learn action-conditioned dynamics in pixel space through video generation (Hafner et al., 2020; Bruce et al., 2024), at a considerably higher planning cost. In all these cases, the temporal model is trained in a separate, action-supervised stage after pretraining. In contrast, our LeVJEPA makes the encoder itself causal during pretraining.

3 Methodology: LeVJEPA

Our goal is to learn visual representations from video with a single network and a single loss. For this, we transfer the LeJEPA paradigm (Balestriero & LeCun, 2025) (i.e., an invariance loss combined with SIGReg as the only mechanism to prevent representation collapse) to video, using a video transformer as backbone (Bardes et al., 2024).

Given a video, we sample a clip of T frames ($T = 16$ during pretraining) and construct $V + 1$ views: one *global* view x_0 at full resolution and V *local* views $x_1, \ldots, x_V$ obtained via aggressive spatial cropping and photometric augmentation. All views cover the same temporal window. Each view $x_v \in \mathbb{R}^{H_v \times W_v \times T \times 3}$ is tokenized into N_v patch tokens of dimension d, the encoder width, of which a random subset of size $L_v = (1 - \rho)N_v$ is retained (Section 3.2). Together with a learnable `[cls]` token, these form the input sequence $S_v \in \mathbb{R}^{(L_v+1)\times d}$, which the shared encoder maps to

the output embeddings $U_v = E_\theta(S_v) \in \mathbb{R}^{(L_v+1)\times d}$. The `[cls]` row $u_v \in \mathbb{R}^d$ provides a clip-level readout, which a projector $g_\phi : \mathbb{R}^d \to \mathbb{R}^K$ maps to the embedding $z_v = g_\phi(u_v) \in \mathbb{R}^K$ on which the loss is computed (with $K = 256$ the dimension of the space that SIGReg regularizes). This projection is necessary because the last encoder layer applies layer normalization (Ba et al., 2016), preventing SIGReg from being optimized effectively in the encoder output space (Kuhn et al., 2026). The projector is discarded after pretraining, and all downstream tasks operate on the encoder outputs U_v rather than on z_v. Figure 1 summarizes the resulting pipeline.

The training objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{inv}} \;+\; \lambda\, \mathcal{L}_{\text{SIGReg}}, \tag{1}$$

where $\mathcal{L}_{\text{inv}}$ encourages the embeddings of the local views to be predictive of the embedding of the global view, and $\mathcal{L}_{\text{SIGReg}}$ constrains the embedding distribution such that collapse is provably prevented (Balestriero & LeCun, 2025). The trade-off weight λ represents the only hyperparameter of the objective. Following Balestriero & LeCun (2025), we set $\lambda = 0.02$ and keep it fixed in all experiments.

3.1 Training Objective

Invariance. The global view is the only view that is not photometrically augmented and covers the largest spatial extent. We therefore use it as the target for all local views and define the invariance term as the mean squared error (MSE)

$$\mathcal{L}_{\text{inv}} \;=\; \frac{1}{V+1} \sum_{v=0}^{V} \|z_0 - z_v\|_2^2 . \tag{2}$$

The target z_0 and the local embeddings z_v are produced by the same encoder in the same forward pass, and gradients flow through both, without any stop-gradient or target network. Minimizing Equation 2 alone would therefore lead to a trivial solution, where the encoder maps every input to the same embedding. In LeVJEPA, this collapse is prevented by SIGReg, described next.

Regularization. SIGReg (Balestriero & LeCun, 2025) constrains the embeddings to follow an isotropic Gaussian. Balestriero & LeCun (2025) show that this distribution minimizes the worst-case downstream probing risk and excludes collapsed solutions by construction. By the Cramér–Wold theorem, the embeddings match $\mathcal{N}(0, I_K)$ if and only if every one-dimensional projection is a standard Gaussian. SIGReg therefore samples M random directions a_m on the unit sphere and penalizes the deviation of each projected batch $\{\langle z_i, a_m\rangle\}_{i=1}^n$ from $\mathcal{N}(0,1)$ via the Epps–Pulley statistic (Epps & Pulley, 1983) (see Appendix A for more details):

$$\mathcal{L}_{\text{SIGReg}} \;=\; \frac{1}{M} \sum_{m=1}^{M} \int \left| \tfrac{1}{n} \textstyle\sum_{i=1}^{n} e^{\mathrm{i}\, t \langle z_i, a_m\rangle} - e^{-t^2/2} \right|^2 e^{-t^2/2}\, dt. \tag{3}$$

3.2 Architecture

The encoder is a Vision Transformer (ViT) adapted to video (Dosovitskiy et al., 2021; Arnab et al., 2021). A clip of T frames is tokenized via a convolutional patch embedding with spatial size 16×16 and temporal size τ; by default $\tau = 1$, i.e., each token corresponds to a patch of a single frame. This results in 3136 tokens for a 224^2 global view and 576 for a 96^2 local view. We compare this choice with the standard temporal aggregation ($\tau = 2$) (Arnab et al., 2021; Tong et al., 2022; Bardes et al., 2024) in Section 4. During pretraining, 95% of the patch tokens of each view ($\rho = 0.95$) are dropped uniformly at random after the patch embedding (Section 4). A learnable `[cls]` token is prepended to the sequence and is never dropped. Only this token receives the training signal, while patch tokens are not supervised directly.

Self-attention is block-causal: patch tokens attend bidirectionally within their frame and to the tokens of previous frames, while the `[cls]` token attends to all tokens but is not attended to. Thus, the representation of a frame only depends on the current and past frames. Positional information is provided by factorized 3D rotary embeddings (Su et al., 2024; Assran et al., 2025), which encode relative position and allow the same encoder to process both view resolutions without interpolation. Apart from the encoder and the projector, no other network is trained. We only keep an EMA of

the encoder weights as evaluation checkpoint (Busbridge et al., 2025), which plays no role during training. Further details are reported in Appendix B.

3.3 Pretraining Data and Evaluation Setup

Pretraining. Following Bardes et al. (2024), we pretrain on a class-balanced 20% subset of K710 (i.e., the union of the Kinetics-400/600/700 training sets (Kay et al., 2017) with validation overlap removed) unless stated otherwise. This keeps the cost of controlled experiments reasonable and, as shown in Section 5, is already sufficient for competitive performance. To study data scaling, we also pretrain a ViT-L/16 on a combination of K710, SSv2 (Goyal et al., 2017), Walking Tours (Venkataramanan et al., 2024), and the PE Video Dataset released with Perception Encoder (Bolya et al., 2026), covering curated action-recognition clips, long uncurated egocentric footage, and web-scale general video.

Evaluation. All encoders, including the baselines, are evaluated frozen, following the attentive probing protocol of Bardes et al. (2024). A cross-attention block with a learnable query pools the output tokens, and a linear classifier is trained on top while the encoder is kept fixed. We use an attentive probe rather than a purely linear one since the pretraining objective does not guarantee that the token representations are linearly separable. We evaluate on IN1K (Russakovsky et al., 2015) for image recognition, Kinetics-400 (K400) (Kay et al., 2017) for appearance-based action recognition, and SSv2 (Goyal et al., 2017) for motion understanding. For K400, given the size of its training set, we use a linear probe on mean-pooled tokens instead. Since this probe is weaker than the attentive one, the reported K400 accuracies are a conservative estimate.

4 Results: What Matters for Efficient Video Pretraining?

We first analyze the design space of LeVJEPA to identify which choices affect the quality of the learned representations. In particular, we focus on three components that are standard in video pretraining: structured masking, temporal patch aggregation, and bidirectional attention. Throughout this section, we pretrain a ViT-B/16 on the 20% subset of K710 (Section 3.3) and report the IN1K top-1 accuracy of a frozen attentive probe, together with SSv2 accuracy where indicated. We start from the tokenization of V-JEPA ($\tau = 2$) with bidirectional attention and modify one design choice at a time, keeping each change in the following experiments.

Finding 1: Uniform random token dropping acts as an augmentation rather than an approximation: accuracy increases as more tokens are dropped, while structured tube masking is harmful.

Increasing the fraction of dropped tokens ρ consistently improves IN1K accuracy, from 33.9% with all tokens to 47.6% at $\rho = 0.95$, while also reducing the cost of each forward pass (Figure 2). Since the clip embedding must be inferred from a sparse, random subset of tokens, dropping acts as an augmentation. Furthermore, going from $\rho = 0.9$ to 0.95 halves the number of processed tokens at no loss in accuracy (47.4% against 47.6%). On SSv2, accuracy decreases for $\rho > 0.3$ at short schedules, but longer training recovers the gap at lower total compute. In contrast, tube dropping, which retains the same spatial locations in every frame, reduces IN1K accuracy from 50.7% to 39.6%. Unlike in masked prediction (Tong et al., 2022), where tube masks prevent copying across frames, here the retained tokens are the only observation of the clip. Tube dropping thus never shows most of the scene, whereas random dropping covers it across frames.

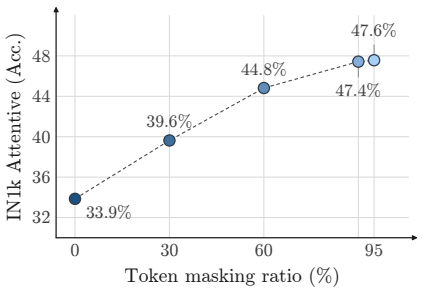

Figure 2: IN1K accuracy as a function of the fraction of dropped tokens ($\tau = 2$).

Increasing the number of local views further improves accuracy, from 47.6% with $V = 4$ to 50.2% with $V = 10$ (see Figure 6 - Appendix D). Since each additional view only adds about 29 tokens, i.e., less than one fifth of the global view, this gain comes at a small cost. We nevertheless keep $V = 4$ by default.

Finding 2: Temporal patch aggregation is not needed: at the same token budget, per-frame tokenization outperforms it, even on motion-centric evaluation.

Video transformers usually aggregate pairs of consecutive frames at the input ($\tau = 2$), halving the number of tokens (Arnab et al., 2021; Tong et al., 2022; Bardes et al., 2024). We compare this choice with per-frame tokenization ($\tau = 1$) at the same cost. During pretraining, the doubled number of tokens of $\tau = 1$ is compensated by a higher dropping ratio ($\rho = 0.95$ against 0.9), so that both encoders process the same number of tokens per view. As shown in Figure 2, these two ratios yield similar accuracy, so this choice does not favor either configuration. At evaluation, where no tokens are dropped, we use 8 frames for $\tau = 1$ and 16 frames for $\tau = 2$ to obtain the same sequence length (for IN1K, images are repeated along time, following Bardes et al. (2024)). As reported in Table 1 (top), per-frame tokenization achieves higher accuracy on both benchmarks. Interestingly, temporal aggregation is usually motivated by the need to capture short-range motion, yet removing it does not hurt accuracy on SSv2, a motion-centric benchmark.

Table 1: Design ablations. *Top:* patch embedding at the same token budget. *Bottom:* attention pattern (τ=1, ρ=0.95, V=4).

	IN1K	SSv2
Patch embedding		
$\tau = 2$, $\rho = 0.90$	47.4	28.8
$\tau = 1$, $\rho = 0.95$	**50.7**	**30.4**
Attention		
Bidirectional	50.7	30.4
Block-causal	**51.2**	**30.7**

Finding 3: Block-causal attention matches bidirectional attention, so temporal ordering can be built into the encoder during pretraining rather than learned afterwards.

We compare bidirectional attention with the block-causal attention described in Section 3.2. Since causal masking removes future frames from the receptive field of each token, a drop in accuracy could be expected. Instead, the block-causal encoder matches its bidirectional counterpart (51.2% against 50.7%, Table 1, bottom). We therefore use block-causal attention in all remaining experiments.

5 Results: Comparison with Prior Work

5.1 Comparison to Video Models

Finding 4: LeVJEPA shifts the accuracy–compute trade-off of video pretraining: it matches the accuracy of video joint-embedding baselines at up to $20.8\times$ less compute, and outperforms them on most benchmarks at the same budget.

Epoch-matched comparison. Comparisons between self-supervised video models are often confounded by differences in pretraining data, schedule, and compute. To remove these factors, we retrain all baselines on the same 20% subsample of K710, using their official implementations and recommended hyperparameters, for the same number of epochs (240) and with the same effective batch size (3072) as our models. All encoders are evaluated with the frozen attentive-probing protocol of Bardes et al. (2024). Since our encoders operate on per-frame tokens, the evaluation sequence length is equalized across methods by adjusting the temporal extent of the input, as described in Section 4, such that all methods are probed on the same number of tokens.

Figure 3 reports IN1K accuracy against total pretraining compute for three encoder sizes. Across all

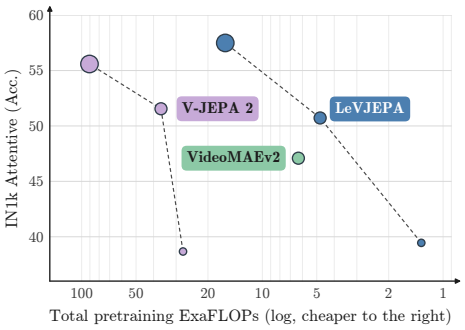

Figure 3: IN1K accuracy against total pretraining compute at matched epochs, for ViT-S, ViT-B, and ViT-L (marker size indicates model size; the horizontal axis is logarithmic and reversed).

sizes, LeVJEPA reaches accuracy comparable to V-JEPA 2 at a fraction of its compute, with the efficiency advantage ranging from $5.6\times$ at ViT-L to $20.8\times$ at ViT-S. At ViT-B, the two methods differ by less than one accuracy point, while LeVJEPA requires 4.8 instead of 36.4 ExaFLOPs. At ViT-L, LeVJEPA outperforms V-JEPA 2 by 1.9 points at $5.6\times$ cost, consuming less than half the compute of the V-JEPA 2 ViT-S. VideoMAEv2 lies in between, with lower accuracy than both methods at an intermediate cost. This efficiency comes directly from the design described in Section 4: full-length token sequences are never processed, and no target encoder or predictor has to be evaluated.

FLOP-matched comparison. The protocol above keeps the number of training samples fixed and lets compute vary. We now fix the total pretraining compute instead, and train all ViT-B encoders with the same FLOP budget. Since LeVJEPA processes each sample at a fraction of the cost of the baselines, the same budget corresponds to a longer schedule of 1085 epochs with $V = 10$ local views. As reported in Table 2, LeVJEPA obtains the highest IN1K accuracy by 7.6 points and the highest K400 accuracy, while remaining within 3.2 points of the strongest baseline on SSv2 (40.4% against 43.6%).

Table 2: FLOP-matched comparison. ViT-B encoders pretrained on the 20% subsample of K710 at equal total FLOPs, evaluated frozen (K400: linear probing).

Method	IN1K	SSv2	K400
VideoMAEv2	53.4	**43.6**	37.4
V-JEPA 2	51.6	42.5	40.7
LeVJEPA	**61.0**	40.4	**44.6**

5.2 Comparison to Image Models

Finding 5: At matched compute on identical data, LeVJEPA approaches image pretraining on appearance and clearly outperforms it on motion, making video a viable substrate for general-purpose visual pretraining.

So far, image pretraining has been the stronger option for appearance-centric transfer: video-pretrained encoders consistently trail their image-pretrained counterparts on image benchmarks, so that video pretraining had to be motivated by motion understanding alone (Venkataramanan et al., 2024; Bardes et al., 2024). We revisit this comparison under matched compute by training DINOv2 (Oquab et al., 2024), with its official implementation, on individual frames from the same videos (11.7M frame samples over 11400 optimizer steps) at the same total FLOPs as the 240-epoch LeVJEPA ViT-B. As reported in Table 3, the image-pretrained encoder retains an advantage of 3.1 points on IN1K, while the video-pretrained one reaches almost twice the accuracy of DINOv2 on SSv2 (30.4% against 16.9%).

Table 3: Comparison to image pretraining at matched compute. ViT-B encoders trained on the same videos; frozen attentive probing.

Method	IN1K	SSv2
DINOv2	**53.8**	16.9
LeVJEPA	50.7	**30.4**

To our knowledge, this is the first FLOP-matched comparison in which video pretraining comes this close to a state-of-the-art image method on appearance, while keeping a clear advantage on motion. The implication goes beyond our method: if the appearance gap can be reduced to a few points at the same compute, then video (which additionally provides temporal structure) becomes the more FLOP-efficient source for general-purpose visual representations.

5.3 Pretraining on Consumer Hardware

The low per-sample cost also changes which hardware suffices for pretraining. We train a ViT-Tiny for 12 hours on a single consumer GPU (RTX 5080, 16 GB) on eight videos of the Walking Tours dataset (Venkataramanan et al., 2024), i.e., unlabeled and uncurated egocentric footage of about 620k frames, from which the model processes roughly 5M clips. IN1K accuracy of the frozen encoder improves from 8.9% with a randomly initialized encoder to 25.2%. Besides the reduced FLOPs, what makes this training practical is the small memory footprint of sparse token sequences: on the same device, LeVJEPA trains with batch size 128 in less than 8 GB, whereas a V-JEPA configuration with the same encoder size, clip length, and resolution saturates the card at batch size

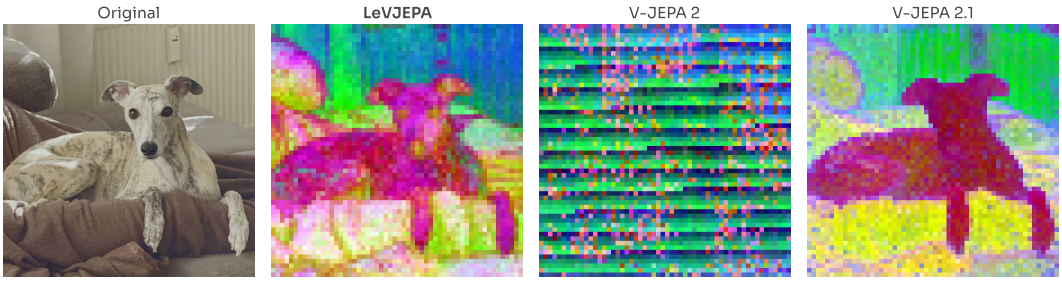

Figure 4: **Patch-token PCA.** The three leading principal components of the patch tokens, visualized as RGB, for the same image. LeVJEPA separates the animal from the furniture and the background, comparably to V-JEPA 2.1, which uses an auxiliary patch-level loss (Mur-Labadia et al., 2026). V-JEPA 2 shows no comparable structure.

28. Thus, with our method, video pretraining is feasible not only at reduced cluster budgets but on commodity hardware.

5.4 Scaling the Pretraining Data

The comparisons above deliberately restrict pretraining to the 20% subsample of K710 to allow controlled, retrained baselines. We now remove this restriction and pretrain a ViT-L/16 for 100 epochs on the combined corpus previously described in Section 3.3. The resulting encoder reaches 69.5% accuracy on IN1K and 55.0% on SSv2 under frozen attentive probing, improving over the ViT-L trained on the 20% subset (Figure 3) by 9.5 points on IN1K within a shorter schedule. Therefore, the method benefits from more data without any change to the objective or its single hyperparameter. Since this corpus is still orders of magnitude smaller than the internet-scale collections of V-JEPA 2 (Assran et al., 2025), the results indicate headroom rather than saturation.

5.5 Emergent Token-Level Structure

The training objective of LeVJEPA supervises a single clip-level token, and the patch tokens receive no loss at any point during training. Nevertheless, these tokens acquire semantically organized representations. In Figure 4, the three leading principal components of the patch tokens, visualized as RGB, group by semantic region and separate the object from its surroundings. V-JEPA 2 does not show a comparable organization, while V-JEPA 2.1 obtains it through an auxiliary patch-level objective (Mur-Labadia et al., 2026), introduced to overcome this limitation of its predecessor. All encoders are evaluated from their publicly released checkpoints.

Figure 5 extends this observation. First, the decomposition is consistent across scenes, i.e., the same components separate object from background in images that differ in content and layout. Second, the cosine similarity between a query patch placed on the object and all other patches is sharply confined to the object, with little response on the background. The representations are therefore not



(a) Cosine similarity (b) Patch-token PCA

Figure 5: **Further patch-token visualizations.** (a) Cosine similarity between a query patch on the object and all other patches. (b) Patch-token PCA for a different scene.

only semantically organized but also spatially precise, which is the property dense prediction tasks such as segmentation, detection, and tracking rely on.

6 Discussion and Conclusion

Our results show that self-supervised learning from video can be considerably simpler and cheaper than currently assumed. A single encoder, trained with an invariance loss and a regularizer whose only hyperparameter is never tuned, reaches accuracy comparable to or better than established video joint-embedding methods at a fraction of their pretraining compute. The target encoder, the predictor, and the schedules these methods rely on for stability are not needed. The analysis of Section 4 extends this simplification to the individual design choices: masking pattern, temporal patch aggregation, and attention topology can be selected for efficiency or for downstream requirements, and the resulting configuration (i.e., a sparse random observation of per-frame tokens with block-causal attention) is at the same time the cheapest and the most accurate we evaluate.

The comparison with image-based pretraining suggests a broader implication. Image pretraining has been the standard source of general-purpose visual representations, with video treated as a specialized complement for motion understanding, to be used when its additional cost is justified. Under our objective, this separation is no longer imposed by efficiency: at equal compute on the same source data, LeVJEPA stays within a few points of DINOv2 on IN1K and almost doubles its accuracy on SSv2. Since video contains the appearance information of its constituent frames while additionally providing temporal structure, these results indicate that video is a viable substrate for general-purpose visual pretraining rather than a specialized one, and that the choice between the two may increasingly depend on data availability rather than on computational cost. The data-scaling experiment of Section 5.4 points in the same direction, with both appearance- and motion-centric accuracy improving on a larger corpus.

Block-causal attention opens a similar perspective for temporal modeling. Current systems obtain causal dynamics by fitting a temporal model on top of a pretrained encoder, which proves to be effective for planning and control (Zhou et al., 2025; Assran et al., 2025; Maes et al., 2026) but separates representation learning from temporal structure. Our results show that the two can be combined: causality can be enforced during pretraining at no measurable cost in downstream accuracy. In this way, the encoder provides per-frame state that respects temporal ordering and extends to incoming frames without re-encoding. This makes a single pretrained encoder a possible starting point for streaming perception and autoregressive world models.

Limitations and future work. Several questions remain open. Motion understanding is the axis where most headroom remains: aggressive token dropping reduces SSv2 accuracy at short schedules, and although longer training recovers the gap within the same compute budget, dropping strategies that preserve temporal correspondences at high sparsity constitute a natural next step of the method. Furthermore, our controlled comparisons use a restricted corpus and models up to ViT-L. The behavior of the objective at the model and data scales of recent video foundation models, as well as the interaction of SIGReg with very large batches, remains to be characterized. In particular, whether the appearance gap to image pretraining persists at internet scale can only be settled by a controlled comparison at that scale, which we do not attempt here. Finally, the patch-level structure reported in Section 5.5 is qualitative. We attribute it to the combination of the invariance loss with aggressive token dropping: since the clip-level embedding has to be inferred from a small random subset of tokens, each retained token must carry information about the clip rather than about its own location alone. This remains a hypothesis, and a quantitative evaluation on dense prediction tasks such as segmentation and tracking is left to future work.

Conclusion. LeVJEPA shows that a single encoder, a single loss, and one fixed hyperparameter are sufficient for competitive self-supervised learning from video. The resulting method is efficient enough to be pretrained on commodity hardware, causal by construction, and improves with more data, without any change to the objective. More broadly, our findings suggest that video is becoming a practical source for general-purpose visual representations, providing appearance and temporal information within a single pretrained encoder. At the same time, the reduced compute and memory requirements lower the barrier to entry: pretraining no longer requires a large cluster, making this line of research accessible to groups with limited resources.

AI USE STATEMENT

We used a large language model as a writing assistant during the preparation of this manuscript. Its use was limited to text we had already drafted: rewriting passages for concision and clarity, checking the consistency of terminology, notation, and cross-references, and identifying typographical errors. The research itself, including the method, the experimental design, the implementation, and the analysis of the results, was carried out by the authors without the assistance of a language model. All text was reviewed by the authors, who take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The training objective is specified in full in Section 3.1, with SIGReg quadrature and distributed-evaluation details in Appendix A; the encoder, projector, and positional-embedding configuration are given in Appendix B; and the frozen attentive- and linear-probing protocols, which follow prior work without modification, in Appendix C. Pretraining corpora and the exact subsampling procedure are described in Section 3.3. All baselines were retrained from their official implementations under the matched protocols of Section 5. The code to reproduce all experiments is included as a zipped folder in the supplementary material, and will be publicly released upon acceptance.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 6816–6826. IEEE, 2021.

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

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

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

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

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.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Thomas W Epps and Lawrence B Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983.

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

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.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

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.

Lukas Kuhn, Giuseppe Serra, Randall Balestriero, and Florian Buettner. Levljepa: End-to-end vision-language pretraining without negatives, 2026. URL https://arxiv.org/abs/2607.00784.

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.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

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.

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(3):211–252, 2015.

Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, 13(2):113–142, 1995.

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

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Shashank Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *International Conference on Learning Representations*, volume 2024, pp. 31952–31973, 2024.

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

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*, pp. 12310–12320. PMLR, 2021.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. 2025 int. conf. *Machine Learning*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

A SIGREG IMPLEMENTATION DETAILS

The integral in Equation 3 is approximated by trapezoidal quadrature over 17 knots $t \in [0, 3]$, and the statistic is computed separately for the embeddings of each view with $M = 1024$ directions per step. Two properties of the formulation are of practical consequence. First, the empirical characteristic function and its gradient are bounded, so the loss is robust to outliers and requires none of the whitening or centering operations of variance–covariance-based regularizers (Bardes et al., 2021). Second, the formulation is compatible with distributed training at negligible cost: the empirical characteristic function is averaged across all workers with a single all-reduce operation before the statistic is computed, such that SIGReg is evaluated on the full global batch rather than on per-device shards. The communication overhead is a single $2 \times M \times 17$ tensor, independent of both the batch size and the embedding dimension.

B ARCHITECTURAL DETAILS

By default, the encoder is a ViT-B/16. The rotary embeddings partition the dimensions of each attention head into three groups, rotated according to the temporal, vertical, and horizontal coordinates of the token, respectively; no absolute positional embeddings are used. The projector is a two-layer multilayer perceptron composed of a linear layer ($d \to 2048$), batch normalization, a GELU non-linearity, and a final linear layer ($2048 \to K$) with $K = 256$; it is discarded after pretraining, and all downstream evaluations operate on the encoder's representations. The Polyak average of the encoder weights uses decay 0.9999, updated every 32 optimizer steps; it receives no forward passes during training and does not appear in the objective, and is therefore distinct from the exponential-moving-average target encoders employed for collapse prevention in prior work (Grill et al., 2020; Bardes et al., 2024).

C EVALUATION DETAILS

Attentive probing. The probe follows Bardes et al. (2024) without modification. A single cross-attention layer attends from one learnable query token to all output tokens of the frozen encoder; the attended output is combined with the query through a residual connection, passed through a two-layer MLP with a GELU non-linearity and layer normalization, and classified by a linear layer. The probe and classifier are trained jointly on the downstream training set while all encoder parameters remain fixed; optimization hyperparameters are taken from Bardes et al. (2024).

Linear probing on Kinetics-400. For Kinetics-400, the output tokens of the frozen encoder are averaged into a single vector and a linear classifier is trained on the pooled representation.

D NUMBER OF LOCAL VIEWS

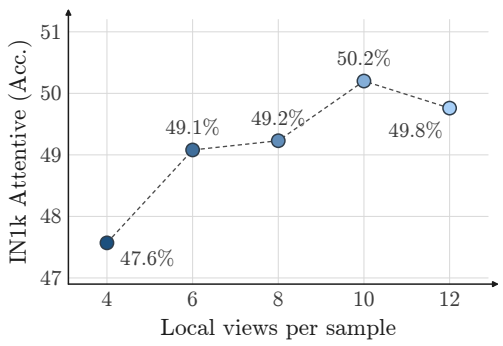

Figure 6: IN1K accuracy as a function of the number of local views V ($\tau = 2$).

Figure 6 reports IN1K accuracy for different numbers of local views V. Accuracy improves from 47.6% at $V = 4$ to 50.2% at $V = 10$ and saturates afterwards (49.8% at $V = 12$). Each additional local view adds one prediction constraint to Equation 2 at a cost of about 29 processed tokens, i.e., less than one fifth of the cost of the global view. Although we keep $V = 4$ in the main comparisons, this shows that additional accuracy can be obtained at a small increase in pretraining cost by enlarging the number of views alone.