arXiv:2303.12001v3 [cs.CV] 2 Oct 2024

ViC-MAE: Self-Supervised Representation Learning from Images and Video with Contrastive Masked Autoencoders

Jefferson Hernandez[1], Ruben Villegas[2], Vicente Ordonez[1]
[1]Rice University, [2]Google DeepMind
{jefehern, vicenteor}@rice.edu, rubville@google.com
Code & Weights: https://github.com/jeffhernandez1995/ViC-MAE

Abstract. We propose ViC-MAE, a model that combines both Masked AutoEncoders (MAE) and contrastive learning. ViC-MAE is trained using a global representation obtained by pooling the local features learned under an MAE reconstruction loss and using this representation under a contrastive objective across images and video frames. We show that visual representations learned under ViC-MAE generalize well to video and image classification tasks. Particularly, ViC-MAE obtains state-of-the-art transfer learning performance from video to images on Imagenet-1k compared to the recently proposed OmniMAE by achieving a top-1 accuracy of 86% (+1.3% absolute improvement) when trained on the same data and 87.1% (+2.4% absolute improvement) when training on extra data. At the same time, ViC-MAE outperforms most other methods on video benchmarks by obtaining 75.9% top-1 accuracy on the challenging Something something-v2 video benchmark. When training on videos and images from diverse datasets, our method maintains a balanced transfer-learning performance between video and image classification benchmarks, coming only as a close second to the best-supervised method.

1 Introduction

Recent advances in self-supervised visual representation learning have markedly improved performance on image and video benchmarks [11,16,40,41]. This success has been mainly driven by two approaches: Joint-embedding methods, which encourage invariance to specific transformations—either contrastive [11, 16, 41] or negative-free [6, 19], and masked image modeling which works by randomly masking out parts of the input and forcing a model to predict the masked parts with a reconstruction loss [4,30,40,92]. These ideas have been successfully applied to both images and video.

Self-supervised techniques for video representation learning have resulted in considerable success, yielding powerful features that perform well across various downstream tasks [30,31,74,92]. Leveraging image-based models to enhance video feature representations has gained widespread adoption, evidenced by significant advancements in robust video representations [2, 55, 58]. The reverse—*video-to-image* transfer learning—has not been as successful. This imbalance underscores a nuanced challenge within multimodal learning, and it is not clear how to

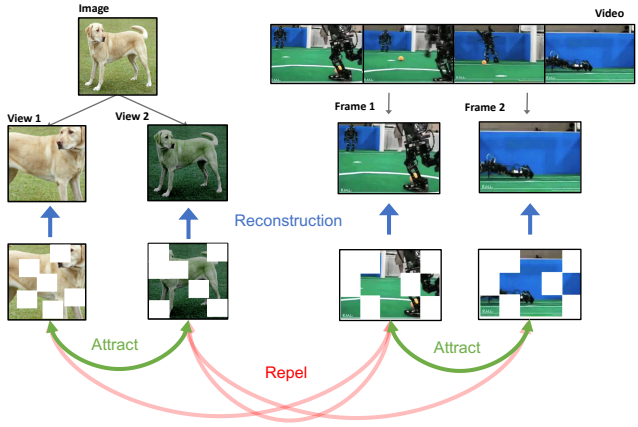


Fig. 1: ViC-MAE operates over video frames and images using masked image modeling at the image and frame level and contrastive learning at the temporal level for videos and under image transformations for images. Our model represents a strong backbone for both image and video tasks.

integrate different modalities. Furthermore, attempts to combine these modalities often result in diminished performance, necessitating tailored adjustments to the underlying architectures or converting one modality (images) into another (repeating images to simulate a video). Learning from video should also yield good image representations since videos naturally contain complex changes in pose, viewpoint, and deformations, among others. These variations can not be simulated through the standard image augmentations used in joint-embedding methods or masked image modeling methods. In this work, we propose a **Vi**sual **C**ontrastive **M**asked **A**uto**E**ncoder (ViC-MAE), a model that learns from both images and video through self-supervision, instead treating short videos as the different views of the same representation, diverging from previous works [33, 34]. On transfer experiments, our model also improves *video-to-image* transfer performance while maintaining performance on video representation learning.

Prior work has successfully leveraged self-supervision for video or images separately using either contrastive learning (*i.e.* Gordon *et al.* [36]), or masked image modeling (*i.e.* Feichtenhofer *et al.* [30]). ViC-MAE seeks to leverage the strength of contrastive learning and masked image modeling and seamlessly incorporate images. While trivially this has been done by repeating the image to simulate a still video, ViC-MAE achieves this in the opposite way, treating frames sampled within short intervals (*e.g.* 1sec) as an additional form of temporal data augmentation. Our method uses contrastive learning to align representations across both time-shifted frames and augmented views, and masked image modeling for single video frames or images to encourage learning local features. Diverging from methods that only use a [CLS] token as a global feature, our model aggregates local features using a global pooling layer followed by a contrastive loss to enhance the representation further. This structure is built upon the foundation of the

Vision Transformer (ViT) architecture [27], which has become a standard for masked image modeling methods.

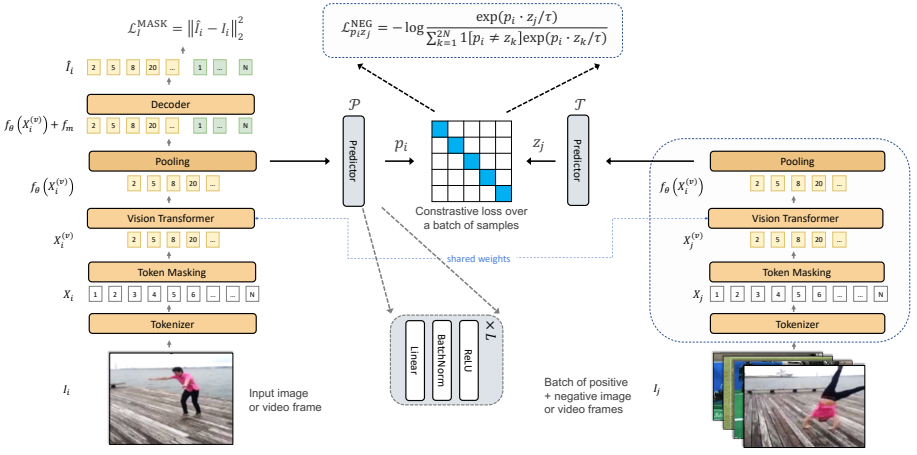


Fig. 2: ViC-MAE inputs two distant frames from a video or two different views of an image within the same batch using a siamese backbone (shared weights), and randomly masks them, before passing them through a ViT model which learns a representation of local features using masked image modeling. A global representation of the video is then constructed by global pooling of the local features learned by the ViT model trained to reconstruct individual patches using an ℓ_2 loss. A standard predictor and a target encoder are used with a contrastive loss. Our use of an aggregation layer before the predictor network aids in avoiding the collapse of the learned global representations.

Closely related to our work is the recently proposed OmniMAE [33] which also aims to be a self-supervised model that can serve as a foundation for image and video downstream tasks. While our experimental evaluations compare ViC-MAE favorably especially when relying on the ViT-L architecture (86% top-1 accuracy on Imagenet vs 84.7%, and 86.8% top-1 accuracy on Kinetics-400 vs 84%), there are also some fundamental differences in the methodology. OmniMAE relies exclusively on masked image modeling and treats images as videos, while ViC-MAE samples frames more sparsely, treating videos within a short time span as the same view. ViC-MAE, leads to reduced training times than video-masked models, while it demands more resources than a basic MAE (which processes 49 visual tokens at a 75% masking rate), it is more efficient (handling 98 tokens at the same rate) than heavier models like OmniMAE or ST-MAE (157 tokens at 90% rate). Surprisingly with these simplications, ViC-MAE works and achieves high performance on video and image tasks, learning effective temporal representations when finetuning on video. Ultimately, we consider our contributions to be orthogonal and could potentially be integrated to achieve further gains.

Our main empirical findings can be summarized as follows: (i) Treating short videos as augmented views, and then finetuning on regular videos or

images yields stronger performance than treating images as videos, while the end models still retain temporal representations, (ii) training with large frame gaps (approx 1.06 seconds) between sampled frames enhances classification performance, providing the kind of strong augmentation that joint-embedding methods typically require, (iii) including negative pairs in training outperforms negative-free sample training,[1] aligning with other methods that have been successful in *video-to-image* evaluations, and (iv) training with strong image transformations as augmentations is necessary for good performance on images.

Our contributions are as follows: (1) We introduce ViC-MAE, which combines contrastive learning with masked image modeling that works on videos and images by treating short videos as temporal augmentations, unlike previous works; (2) When ViC-MAE is trained only on videos, we achieve state-of-the-art *video-to-image* transfer learning performance on the ImageNet-1K benchmark and state-of-the-art self-supervised performance for video classification on SSv2 [38]; and (3) We demonstrate that ViC-MAE achieves superior transfer learning performance across a wide spectrum of downstream image and video classification tasks, outperforming baselines trained only with masked image modeling. Our source code and model checkpoints are available here.

2 Related Work

Our work is related to various self-supervised learning strategies focusing on video and image data, especially in enhancing image representation through video.

Self-supervised Video Learning. Self-supervised learning exploits temporal information in videos to learn representations aiming to surpass those from static images by designing pretext tasks that use intrinsic video properties such as frame continuity [26, 61, 64, 79, 85, 86], alongside with object tracking [1, 72, 90, 91]. Contrastive learning approaches on video learn by distinguishing training instances using video temporality [6, 19, 36, 71, 93, 95]. Recently, Masked Image Modeling (MIM) has used video for pre-training either using the standard design [40] or an asymmetrical siamese design that predicts future masked frames conditioned on present unmasked frames [39]; aiding in transfer learning for various tasks [30,83,92]. Our approach uniquely integrates contrastive learning and masked image modeling into a single pre-training framework suitable for image and video downstream applications.

Learning *video-to-image* representations. Several previous models trained only on images have demonstrated remarkable *image-to-video* adaptation [2,55,58]. However, static images lack the dynamism inherent to videos, missing motion cues and camera view changes. In principle, this undermines image-based models for video applications. Recent work has leveraged video data to learn robust image representations to mitigate this. For instance, VINCE [36] shows that natural augmentations found in videos could outperform synthetic augmentations. VFS [95] uses temporal relationships to improve results on static image tasks.

[1] See supplemental material for an evaluation of what we tried and did not work when combining negative-free methods with masked image modeling

CRW [93] employs cycle consistency for inter-video image mapping, allowing for learning frame correspondences. ST-MAE [30] shows that video-oriented masked image modeling can benefit image-centric tasks. VITO [71] develops a technique for video dataset curation to bridge the domain gap between video and images.

Learning general representations from video and images. Research has progressed in learning from video and images, adopting supervised or unsupervised approaches. The recently proposed TubeViT [73] uses sparse video tubes for creating visual tokens across images and video. OMNIVORE [34] employs a universal encoder for multiple modalities with specific heads for each task. PolyViT [56] additionally trains with audio data, using balanced task-training schedules. Expanding on the data modalities, ImageBind [32] incorporates audio, text, and various sensor data, with tailored loss functions and input sequences to leverage available paired data effectively. In self-supervised learning, BEVT [88] adopts a BERT-like approach for video, finding benefits in joint pre-training with images. OmniMAE [33] proposes masked autoencoding for joint training with video and images. OmniVec [80] extends the datasets using in OMNIVORE, creates new task training policies, and adds masked autoencoding as an auxiliary task to learn from multiple modalities. ViC-MAE learns from video and image datasets without supervision by combining masked image modeling and contrastive learning.

Combining contrastive methods with masked image modeling. Contrastive learning combined with masked image modeling has been recently investigated. MSN [3] combines masking and augmentations for efficient contrastive learning, using entropy maximization instead of pixel reconstruction to avoid representational collapse, achieving notable few-shot classification performance on ImageNet-1k. CAN [65] uses a framework that combines contrastive and masked modeling, employing a contrastive task on the representations from unmasked patches and a reconstruction plus denoising task on visible patches. C-MAE [46] uses a Siamese network design comprising an online encoder for masked inputs and a momentum encoder for full views, enhancing the discrimination power of masked autoencoders which usually lag in linear or KNN evaluations. C-MAE-V [62] adapts C-MAE to video, showing improvements on Kinetics-400 and Something Something-v2. MAE-CT [51] leverages a two-step approach with an initial masked modeling phase followed by contrastive tuning on the top layers, improving linear classification on masked image modeling-trained models. Our ViC-MAE sets itself apart by effectively learning from images and videos within a unified training approach, avoiding the representational collapse seen in C-MAE through a novel pooling layer and utilizing dual image crops from data augmentations or different video frames to improve modality learning performance.

3 Method

We propose ViC-MAE for feature learning on video and images, which works using contrastive learning at the temporal level (or augmentations on images) and masked image modeling at the image level.

3.1 Background

We provide below some background terminology and review of closely related methods that we build upon.

Masked image modeling. This approach provides a way to learn visual representations in a self supervised manner. These methods learn representations by first masking out parts of the input and then training a model to fill in the blanks using a simple reconstruction loss. To do this, these methods rely on an encoder f_θ that takes the non-masked input and learns a representation x, such that a decoder d_ϕ can reconstruct the masked part of the input. More formally, let x be the representation learned by the encoder for masked image I with mask M such that $f_\theta(I \odot M)$. A decoder d is then applied to obtain the first loss over masked and unmasked tokens $d_\phi(x)$. This defines the following reconstruction loss which is only computed over masked tokens:

$$\mathcal{L}_I^{\text{MASK}} = \left\| d_\phi(f_\theta(I \odot M)) \odot (1 - M) - I \odot (1 - M) \right\|_2 . \tag{1}$$

Contrastive learning. In common image-level contrastive methods, learning with negatives is achieved by pushing the representation of the positive pairs (different augmented views of the same image) to be close to each other while pulling the representation of negative pairs further apart. More formally, let I and I' be two augmented views of the same image. Contrastive learning uses a siamese network with a prediction encoder $\mathcal{P}$ and a target encoder $\mathcal{T}$ [16,95]. The output of these networks are ℓ_2-normalized: $p = \mathcal{P}(I)/\|\mathcal{P}(I)\|_2$, and $z = \mathcal{T}(I')/\|\mathcal{T}(I')\|_2$. Given a positive pair from a minibatch of size N, the other $2(N-1)$ examples are treated as negative examples. The objective then is to minimize the InfoNCE loss [68]. When learning with negatives, $\mathcal{P}$ and $\mathcal{T}$ typically share the same architecture and model parameters.

3.2 ViC-MAE

We propose a novel approach for learning representations by applying masking image modeling at the individual image level, paired with image-level similarity using either sampled frames or augmented images. Unlike previous methods that inefficiently replicate images to mimic video input, thereby utilizing more computational resources, our methodology treats short video segments as augmented instances of a single view. This perspective not only enhances the efficiency of the learned representations but also significantly broadens the applicability of our model. ViC-MAE offers a versatile "plug and play" solution for image-based tasks. Furthermore, our model can easily be fine-tuned for video tasks and adapted to videos of varying sizes, unlike the traditional 16 frames. Figure 2 shows an overview of our model.

Given a video with T frames $\{I_1, I_2, \cdots, I_T\}$, we sample two frames I_i, I_j as a positive pair input during one training step. We augment single images when they appear in a batch. Notice that our model sees a batch comprising frames and images. After an input image tokenizer layer, we obtain a set of patch-level token

representations of X_i and X_j for each frame. Then, we apply token masking by generating a different random mask M_i and M_j and apply them to both of the corresponding input frames to obtain a subset of input visible tokens $X_i^{(v)}$ and $X_j^{(v)}$. These visible token sets are then forwarded to a ViT encoder, which computes a set of representations $f_\theta(X_i^{(v)})$ and $f_\theta(X_j^{(v)})$ respectively. Finally, for the first image, we compute $\hat{I}_i = d_\phi(f_\theta(X_i^{(v)} + f_m))$ where we have added a mask token f_m to let the decoder know which patches were masked and allows to predict patch-shaped outputs through $\hat{I}_i$. These output patches are then trained to minimize the ℓ_2 loss with the true patches in the input image:

$$\mathcal{L}_i^{\text{MASK}} = \|\hat{I}_i - I_i\|_2^2. \tag{2}$$

To apply contrastive pre-training we use a separate prediction branch in the network by applying a global pooling operator Ω over the output representations $f_\theta(X_i^{(v)})$ from the main branch and $f_\theta(X_j^{(v)})$ from the siamese copy of the network. This step simplifies the formulation of our method and avoids using additional losses or the `gradient-stop` operator as in SimSiam [19] to avoid feature representation collapse since the pooled features can not default to the zero vector as they also are being trained to reconstruct patches. We experiment using various aggregation methods, including *mean* pooling, *max* pooling, and *generalized mean* (GeM) pooling [75].

These global representations are then forwarded to a predictor encoder $\mathcal{P}$ and a target encoder $\mathcal{T}$ to obtain frame representations:

$$p_i \triangleq \mathcal{P}(\Omega(f_\theta(X_i^{(v)})))/\|\mathcal{P}(\Omega(f_\theta(X_i^{(v)})))\|_2,$$

and

$$z_j \triangleq \mathcal{T}(\Omega(f_\theta(X_j^{(v)})))/\|\mathcal{T}(\Omega(f_\theta(X_j^{(v)})))\|_2$$

respectively. The predictor network $\mathcal{P}$ and target network $\mathcal{T}$ are symmetrical and we use standard blocks designed for contrastive learning [6, 16, 19]. These blocks consist of a Linear $\rightarrow$ BatchNorm1d $\rightarrow$ ReLU block repeated 2 times. From these representations, we apply the InfoNCE contrastive learning loss as follows:

$$\mathcal{L}_{p_i,z_j}^{\text{NEG}} = -\log \frac{\exp(p_i \cdot z_j/\tau)}{\sum_{k=1}^{2N} \mathbb{1}[p_i \neq z_k]\exp(p_i \cdot z_k/\tau)}, \tag{3}$$

where the denominator includes a set of negative pairs with representations z_k computed for frames from other videos, the same video but at a time longer than the selected time shift and images in the same batch, $\mathbb{1}[p_i \neq z_k] \in \{0, 1\}$ is an indicator function evaluating to 1 when $p_I \neq z_k$ and τ denotes a temperature parameter.

The final loss is $\mathcal{L} = \mathcal{L}^{\text{MASK}} + \lambda\mathcal{L}^{\text{NEG}}$, where λ is a hyperparameter controlling the relative influence of both losses. In practice, we use a schedule to gradually introduce the contrastive loss and let the model learn good local features at the beginning of training.

Table 1: Transfer learning results from video and image pre-training to various datasets using the ViT/L-16 backbone. The pre-training data is a video dataset (MiT, K600, K700, or K400) and/or image dataset (IN1K). All self-supervised methods are evaluated end-to-end with supervised finetuning on IN1K, Kinetics-400, Places365, and SSv2. Best results are in bold. Results of MAE, ST-MAE, and VideoMAE for out-of-domain data were taken from Girdhar *et al.* [33].

	Method	Arch.	Pre-training Data	In-Domain		Out-of-Domain	
				IN1K	K400	Places-365	SSv2
Supervised	ViT [27] *ICML'20*	ViT-B	IN1K	82.3	68.5	57.0	61.8
	ViT [27] *ICML'20*	ViT-L	IN1K	82.6	78.6	58.9	66.2
	COVeR [97] *arXiv'21*	TimeSFormer-SR	JFT-3B+ K400+ MiT + IN1K	86.6	87.2	-	70.9
	OMNIVORE [34] *CVPR'22*	ViT-B	IN1K + K400 + SUN RGB-D	84.0	83.3	59.2	68.3
	OMNIVORE [34] *CVPR'22*	ViT-L	IN1K + K400 + SUN RGB-D	86.0	84.1	–	–
	TubeViT [73] *CVPR'23*	ViT-B	K400 + IN1K	81.4	88.6	–	–
	TubeViT [73] *CVPR'23*	ViT-L	K400 + IN1K	–	90.2	–	76.1
Self-Supervised	MAE [40] *CVPR'22*	ViT-B	IN1K	83.4	–	57.9	59.6
	MAE [40] *CVPR'22*	ViT-L	IN1K	85.5	82.3	59.4	57.7
	ST-MAE [30] *NeurIPS'22*	ViT-B	K400	81.3	81.3	57.4	69.3
	ST-MAE [30] *NeurIPS'22*	ViT-L	K400	81.7	84.8	58.1	73.2
	VideoMAE [83] *NeurIPS'22*	ViT-B	K400	81.1	80.0	–	69.6
	VideoMAE [83] *NeurIPS'22*	ViT-L	K400	–	85.2	–	74.3
	OmniMAE [33] *CVPR'23*	ViT-B	K400 + IN1K	82.8	80.8	58.5	69.0
	OmniMAE [33] *CVPR'23*	ViT-L	K400 + IN1K	84.7	84.0	59.4	73.4
	ViC-MAE	ViT-L	K400	85.0	85.1	59.5	73.7
	ViC-MAE	ViT-L	MiT	85.3	84.9	59.7	73.8
	ViC-MAE	ViT-B	K400 + IN1K	83.0	80.8	58.6	69.5
	ViC-MAE	ViT-L	K400 + IN1K	86.0	86.8	60.0	75.0
	ViC-MAE	ViT-B	K710+ MiT + IN1K	83.8	80.9	59.1	69.8
	ViC-MAE	ViT-L	K710 + MiT + IN1K	**87.1**	**87.8**	**60.7**	**75.9**

4 Experiment Settings

We perform experiments to demonstrate the fine-tuning performance of our method on ImageNet-1k and other image recognition datasets. We also evaluate our method on the Kinetics-400 dataset [47] and Something Something-v2 [38] for action recognition to show that our model is able to maintain performance on video benchmarks. Full details are in the supplemental material.

Architecture. We use the standard Vision Transformer (ViT) architecture [27] and conduct experiments fairly across benchmarks and methods using the ViT-B/16 and ViT-L/16 configurations. For masked image modeling, we use a small decoder as proposed by He *et al.* [40]. Finetunig on images requires no changes since this resembles the pre-training configuration. Finetuning on videos is as follows: we initialize the temporal tokenizer by replicating the spatial tokens along the temporal dimension scaled by the length of the video, similarly, we initialize the MHA parameters by replicating them but skip the scaling for them. We use the standard of finetuning on videos of 16 frames, skipping 4.

Pre-Training. We adopt Moments in Time [66], Kinetics-400 [47], and ImageNet-1k [25] as our main datasets for self supervised pre-training. They consist of

∼1000K and ∼300K videos of varied length respectively, and ∼1.2M images for Imagenet-1k. We sample frames from these videos using distant sampling, which consists of splitting the video into non-overlapping sections and sampling one frame from each section. Frames are resized to a 224 pixel size, horizontal flipping, and random cropping with a scale range of $[0.5, 1]$, as the only data augmentation transformations on video data. Random cropping (with flip and resize), color distortions, and Gaussian blurring are used for the image modality. For our largest training run, we combine the training sets of Kinetics-400 [47], Kinetics-600 [12], and Kinetics-700 [13], with duplicates removed based on YouTube IDs. We also exclude K400 videos used for evaluation from training to avoid leakage. This process results in a unique, diverse dataset of ∼665K samples, which we label K710, following [87].

Settings. We follow previously used configurations for pre-training [30, 40]. We use the AdamW optimizer with a batch size of 512 per device. We evaluate the pre-training quality by end-to-end finetuning. When evaluating on video datasets we follow the common practice of multi-view testing: taking K temporal clips ($K = 7$ on Kinetics) and for each clip taking 3 spatial views to cover the spatial axis (this is denoted as $K \times 3$). The final prediction is the average of all views.

5 Results and Ablations

We first perform experiments to analyze the different elements of the ViC-MAE framework. All the experiments are under the *learning with negative pairs* setting using mean pooling over the ViT features. Linear evaluation and end-to-end finetuning runs are done over 100 epochs for ImageNet-1k, see supplemental material for more details. For our ablations, we restrict ourselves to the ViT-B/16 architecture pre-trained over 400 epochs unless specified otherwise.

5.1 Main result

Our main result evaluates ViC-MAE on two in-domain datasets that were used during training for most experiments: ImageNet-1K (images) and Kinetics-400 (video), and two out-of-domain datasets that no methods used during training: Places-365 [99] (images) and Something-something-v2 (video). Table 1 shows our complete set of results including comparisons with the state-of-the-art on both supervised representation learning (typically using classification losses), and self-supervised representation learning (mostly using masked image modeling). We consider mostly recent methods building on visual transformers as the most recent TubeViT [73] which relies on this type of architecture.[2]

Our most advanced version of ViC-MAE trained on five datasets (Kinetics-400, Kinetics-600, Kinetics-700, Moments in Time, and Imagenet-1k) using the ViT-Large architecture performs the best across all metrics on all datasets

[2]Previous methods also use different backbones [36, 93, 95] *i.e.* ResNet-50. They obtain 54.5%, 33.8%, and 55.6% top-1 accuracies on linear evaluation on ImageNet-1k. Since those works do not use the same setting, we do not include them here.

Table 2: Comparison of transfer learning performance of our approach with supervised baselines across 8 natural image classification datasets. All results correspond to linear evaluation. Best results are shown in bold. ‡MAE trained on MiT and K400 randomly sample a frame from the video to compute a reconstruction loss; these models are trained and evaluated by us. See supplemental material for more evaluation of transfer learning performance.

	Model	Pre-train.	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	VOC2007	DTD	Caltech101
ViT/B-16	MAE [40] ‡	K400	74.54	94.86	79.49	46.51	64.33	83.07	78.01	93.28
	MAE [40] ‡	MiT	76.23	94.47	79.50	47.98	65.32	83.46	78.21	93.08
	ViC-MAE (ours)	K400	76.56	93.64	78.80	47.56	64.75	83.74	78.53	92.27
	ViC-MAE (ours)	MiT	**77.39**	**94.92**	**79.88**	**48.21**	**65.64**	**84.77**	**79.27**	**93.53**
ViT/L-16	MAE [40]	IN1K	77.5	95.0	82.9	49.8	63.2	83.3	74.5	94.8
	OmniMAE [33]	SSv2+IN1K	76.2	94.2	82.2	50.1	62.6	82.7	73.9	94.4
	ViC-MAE (ours)	IN1K+K400	81.9	95.6	85.4	52.8	67.3	84.2	76.8	94.9
	ViC-MAE (ours)	K710+MiT+IN1K	**82.9**	**96.8**	**86.5**	**53.5**	**68.1**	**85.3**	**77.8**	**96.1**

compared to all previous self-supervised representation learning methods and even outperforms the supervised base model OMNIVORE [34] on Imagenet-1k with a top-1 accuracy of 87.1% vs 86%. As well as, COVeR [97] a model trained on a similar data mix, except that it uses more images, COVeR gets 86.6% vs 87.1% on Imagenet-1k. ViC-MAE also comes a close second to other supervised methods and roughly matches the performance of TubeViT [73] which obtains 76.1% top-1 accuracy on Something something-v2 compared to our 75.9% top-1 accuracy. When compared to the current self-supervised state-of-the-art OmniMAE using the same ViT-Large architecture and the same datasets for pre-training (Kinetics-400 and Imagenet-1k), ViC-MAE also outperforms OmniMAE in all benchmarks (Imagenet: 86% *vs*. 84.7%, Kinetics-400: 86.8% *vs*. 84%, Places-365: 60% *vs*. 59.4% and SSv2: 75% vs 73.4%).

Another important result is *video-to-image transfer*, where the model is only trained on video but its performance is tested on downstream image tasks. Table 1 shows that when ViC-MAE is trained on the Moments in Time dataset [66], it achieves the best top-1 accuracy of 85.3% for any self-supervised backbone model trained only on video. These results highlight the closing gap in building robust representations that can work seamlessly across image and video tasks.

5.2 Comparison with other contrastive masked autoencoders.

Combining MAE with joint-embedding methods is non-trivial. In our first attempts, we used the [CLS] token as the representation and applied negative free methods such as VicReg [6], and SimSiam [19] with limited success (See supplemental material). When combined with contrastive methods, we found it best to use a pooling operation over the ViT features similar to CAN [65], as we find worse performance when the [CLS] token is used, like in C-MAE [46]. The original MAE [40] is known to have poor linear evaluation performance, obtaining 68% in IN1K linear evaluation when pre-trained on IN1K [40, 51]. On the contrary, SimCLR [17] a model trained only using contrastive learning on IN1K

Table 3: ViC-MAE ablation experiments with ViT/B-16. We present linear evaluation results on the ImageNet-1K dataset.

(a) Ablation on frame separation. 0: sample same frame, D: distant sampling, and > 0 continuous sampling.

Frame separation	ImageNet-1K	
	Top-1	Top-5
0	63.25	83.34
2	64.47	84.31
4	65.25	84.64
8	65.89	84.91
D	**67.66**	**86.22**

(b) Ablation on pooling type. The hyperparameter λ is set to 0.025 and introduced using a schedule.

Pooling type	Top-1	Top-5
GeM	66.92	85.50
max	67.01	85.59
mean	**67.66**	**86.22**

(c) Ablation on different augmentations.We use a combination of different color and spatial augs.

Color Augm.	Spatial Augm.	ImageNet-1K	
		Top-1	Top-5
✓		65.40	84.03
	✓	66.03	85.01
✓	✓	**67.66**	**86.22**

achieves 73.5%. Several works have tried to address this by combining contrastive learning with masked image modeling to get the best of both worlds. CAN [65], C-MAE [46] and MAE-CT [51] obtain linear evaluation accuracies of 74.0%, 73.9, 73.4%, respectively when trained on IN1K while ViC-MAE obtains 74.0% trained only on IN1K using ViT/B-16 pre-trained for 800 epochs to make the comparison fair. When using the K400 and IN1K datasets together for pre-training, we get 73.6%, but we highlight that ViC-MAE can now maintain good performance in videos and images using the same pre-trained model.

5.3 Transfer Learning Experiments

In this section, we evaluate our pre-trained models from Table 1 for transfer learning on downstream tasks.

Video-to-image transfer learning performance. We evaluate transfer learning performance of ViC-MAE across a diverse array of 12 downstream image classification tasks [7, 9, 21, 29, 48, 49, 63, 67, 70, 94]. (Due to space constraints, we have shown the six most significant ones. See supplemental material for the full table.) Table 2 shows the results of four models based on a ViT/B backbone. We perform linear evaluation. We train two models using two video datasets. The first model is a baseline MAE model pre-trained on randomly sampled frames from videos on the Moments in Time and Kinetics-400 datasets. The second model is our full ViC-MAE model pre-trained on each of the same two datasets. Our model significantly outperforms the other baselines on 9 out of 12 datasets, whereas the MAE trained on Kinetics is superior on only 3 (i.e. Cars, Aircraft, and Pets). When scaling the size of our models, we see that ViC-MAE surpasses all models, including OmniMAE [33] trained on SSv2+IN1K[3]

Object detection and segmentation. We finetune Mask R-CNN [42] end-to-end on the COCO dataset. We adapted the ViT backbone to be used with the FPN, following the recipe outlined in Li *et al*. [54]. We apply this approach

[3]These are the only publicly available checkpoints of OmniMAE

Table 4: COCO object detection and segmentation using a ViT-B Mask R-CNN baseline. All entries use data without labels.

Method	pre-train data	AP_{Box}	AP_{Mask}
MAE [40]	IN1K	50.3	44.9
C-MAE [46]	IN1K	52.4	46.5
ViC-MAE	IN1K+K400	52.5	46.5
ViC-MAE	IN1K+K710+MiT	**53.2**	**46.9**

to ViC-MAE and take the other results from their respective paper. See Table 4. Compared with previous methods, ViC-MAE outperforms other approaches under the same configurations. Specifically, when utilizing the combined IN1K+K400 dataset, ViC-MAE achieves a box AP of 52.5 and a mask AP of 46.5, slightly improving over C-MAE, which stands at 52.4 for box AP and 46.5 for mask AP. More notably, with the expanded dataset of IN1K+K710+MiT, ViC-MAE significantly advances the state-of-the-art, achieving the highest reported scores of 53.2 for box AP and 46.9 for mask AP.

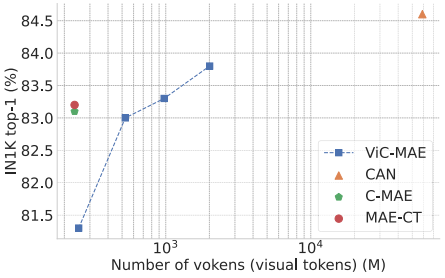
5.4 Ablations

We investigate the effect of scaling the data used to train ViC-MAE, the effect of the ratio of image to videos in pre-training, our choice of frame separation, the choice of pooling operator, and the choice of data augmentations. An extra ablation probing the temporal representation learning of our method can be found in the supplemental material.

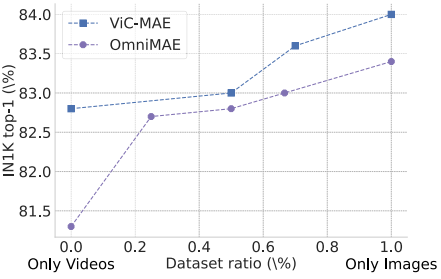
Influence of pre-training data. We perform an ablation study to the effect of scaling the data points seen by the model. The pre-training data includes Kinectis-400, ImageNet-1K, Kinectis-600 + Kinectis-700, and the Moments in Time datasets added in that order. We pre-train a ViT/B-16 using ViC-MAE for 400 epochs. As illustrated in Figure 3a, as we progressively increase the dataset size, our ViC-MAE, shows a steady increase in IN1K top-1 accuracy. This is remarkable when compared to CAN [65], pre-trained on the JFT-300M dataset for 800 epochs that only reaches an accuracy of 84.4%. This shows that our model, when supplied with only about 1.5% of the data that CAN was trained on (4.25M vs. 300M), can achieve comparable accuracy levels.

Contrastive vs Masking-only pre-training We perform an ablation study to the effect of varying the ratio of images to video in the dataset by replicating the entire dataset; notice that the number of training updates changes when doing this. The pre-training data includes Kinectis-400 and ImageNet-1K. We pre-train a ViT/B-16 using ViC-MAE for 400 epochs. As illustrated in Figure 3b, as we progressively increase the ratio of images to videos, our ViC-MAE, surpasses the OmniMAE model [95], meaning that contrastive plus masking pre-training is better able to use image and video data than masking-only pre-training.

Frame separation. We aim to explore the effect of frame separation on model performance. We follow the two methods of sampling frames from Xu et.al [95].



(a) ViC-MAE ViT/B-16 finetuned on IN1K for 100 epochs, compared with CAN pre-trained on JFT-300M, C-MAE, and MAE-CT pre-trained on ImageNet-1K. We increase the amount of data points by adding more video datasets. We can see that our model reaches similar accuracy with $\approx$ 4.25M data points compared to the 300M of the JFT-300M dataset.



(b) ViC-MAE using the ViT/B-16 architecture finetuned on IN1K for 100 epochs, compared with OmniMAE. We vary the ratio of images vs video in the dataset, from no images to only images. We can see that ViC-MAE can better utilize the videos and images on the dataset compared to masking-only pre-training.

Fig. 3: Additional comparisons with the state-of-the-art and recently proposed methods.

Results are shown in Table 3a. The first approach, *Continuous sampling*, involves selecting a start index i and sampling a frame within $(i, i+\delta]$, where δ represents the frame separation, with a separation of 0 meaning identical frames for predictor and target networks. The second, *Distant sampling*, divides the video into n equal intervals, corresponding to the number of frames for contrastive learning, and randomly selects one frame from each interval.

In our experiment, we observe that increasing the frame separation when using *continuous sampling* increases model performance. We observe the best performance using *distant sampling* with $n = 2$ (labeled D in Table 3a). We posit that further increasing frame separation offers potentially stronger augmentations. In the following experiments, we only use strong spatial augmentations combined with distant frame sampling.

Pooling type. We test which operator Ω used to aggregate local features performs best at producing global features. We report our results in Table 3b. We try common types of pooling (*mean, max*) as well as, *generalized mean pooling*. We found *mean* to be more effective in creating a global representation for video, and we use it for all other experiments.

Adding strong augmentations to video frames In our ablation study, we investigated the necessity of strong color augmentations for video frames during joint training with the target encoder, as commonly applied to images. The findings, detailed in Table 3c, indicate a performance decrease of over 2% in linear evaluation on the Imagenet dataset when applying solely color augmentations without spatial adjustments. Interestingly, employing color and spatial augmentations does not outperform strong spatial augmentations alone. This diverges from prior approaches that rely heavily on color augmentations for effective contrastive learning, suggesting that the inherent temporal variations in video frames may suffice. However, for image datasets, the combination of strong color and spatial augmentations remains necessary.

5.5 Limitations

Our proposed model is able to learn representations from video and image data that transfer to several downstream tasks and surpasses previous models on the same set-up. Given a similar setup, ViC-MAE matches prior results on Kinetics-400, trailing only to supervised models such as TubeViT [73] by 7.1% on ViT/B-16 and 2.4% on ViT/L-16, and MVT [96] slightly on ViT/B-16 but surpasses it by 3.5% on ViT/L-16. It also exceeds MViTv1 [28], TimeSformer [8], and ViViT [2] by margins up to 7.3% on ViT/L-16. Compared to self-supervised models, ViC-MAE falls behind MaskFeat [92] by 0.7% on ViT/B-16 but excels on ViT/L-16 by 3.5%. Our model surpasses V-JEPA [5] by margins up to 1.1% but falls behind VideMAEV2 [87] by 0.8% on ViT/L-16. It is slightly outperformed by DINO [11] and more substantially by models using extra text data or larger image datasets, such as UMT [53], MVD [89], and UniFormerV2 [52], by up to 4.2% on ViT/B-16 and 2.8% on ViT/L-16. Future work could consider leveraging additional weak supervision through other modalities such as text [10, 43, 76], audio [77, 82], 3D geometry [78] or automatically generated data [14, 44].

When compared against state-of-the-art ImageNet-pretrained models with comparable computational resources, video-based models, including ours, typically fall short. However, including image and video modalities shows promise in boosting performance. Against models using masked image modeling and contrastive learning, ViC-MAE modestly surpasses MAE [40] by 1.6%, MaskFeat [92] by 1.4% and iBOT [100] by 0.5% with the ViT/L-16 architecture. It also edges out MoCov3 [18] and BeiT [4] by 3% and 1.9% respectively on the same architecture. Yet, it lags behind DINOv2 [69] by 1.2% for ViT/L-16. When compared to supervised models using additional image data, such as DeiT-III [84] and SwinV2 [57] and the distilled ViTs' from [24], our model shows a lag behind of 0.8%, 1.2% and 2.5% respectively on ViT/L-16. These results show that the gap from models pre-trained purely on video still exists, but we believe ViC-MAE pre-trained on image and video data is a step forward in closing that gap.

6 Conclusion

In this work, we introduce ViC-MAE, a method that allows to use unlabeled videos and images to learn useful representation for image recognition tasks. We achieve this by randomly sampling frames from a video or creating two augmented views of an image and using contrastive learning to pull together inputs from the same video and push apart inputs from different videos, likewise, we also use masked image modeling on each input to learn good local features of the scene presented in each input. The main contribution of our work is showing that it is possible to combine masked image modeling and contrastive learning by pooling the local representations of the MAE prediction heads into a global representation used for contrastive learning. The design choices that we have taken when designing ViC-MAE show that our work is easily extensible in various ways. For example, improvements in contrastive learning for images can be directly adapted into our framework. Likewise, pixel reconstruction can be replaced by features important for video representation such as object correspondences or optical flow.

Acknowledgements

The authors would like to thank Google Cloud and the CURe program from Google Research for partially providing funding for this research effort. We are also thankful for support from the Department of Computer Science at Rice University, the National Science Foundation through NSF CAREER Award #2201710, and the Ken Kennedy Institute at Rice University. We also thank anonymous reviewers for their feedback and encouragement.

References

1. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: Proceedings of the IEEE international conference on computer vision. pp. 37–45 (2015)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning. In: European Conference on Computer Vision. pp. 456–473. Springer (2022)
4. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (2021)
5. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
6. Bardes, A., Ponce, J., Lecun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: ICLR 2022-International Conference on Learning Representations (2022)
7. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2011–2018 (2014)
8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: International Conference on Machine Learning. pp. 813–824. PMLR (2021)
9. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014)
10. Cai, M., Liu, H., Mustikovela, S.K., Meyer, G.P., Chai, Y., Park, D., Lee, Y.J.: Vip-llava: Making large multimodal models understand arbitrary visual prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12914–12923 (June 2024)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
12. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018)

13. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
14. Cascante-Bonilla, P., Shehada, K., Smith, J.S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., Karlinsky, L.: Going beyond nouns with vision & language models using synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20155–20165 (October 2023)
15. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International conference on machine learning. pp. 1691–1703. PMLR (2020)
16. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
17. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
18. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
19. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
20. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
21. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
22. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations (2019)
23. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
24. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning. pp. 7480–7512. PMLR (2023)
25. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
26. Diba, A., Sharma, V., Gool, L.V., Stiefelhagen, R.: Dynamonet: Dynamic action and motion network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6192–6201 (2019)
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
28. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6824–6835 (2021)

29. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
30. Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. Neural Information Processing Systems (NeurIPS) (2022)
31. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3299–3309 (2021)
32. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
33. Girdhar, R., El-Nouby, A., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Omnimae: Single model masked pretraining on images and videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10406–10417 (2023)
34. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A single model for many visual modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16102–16112 (2022)
35. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
36. Gordon, D., Ehsani, K., Fox, D., Farhadi, A.: Watching the world go by: Representation learning from unlabeled videos. arXiv preprint arXiv:2003.07990 (2020)
37. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
38. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
39. Gupta, A., Wu, J., Deng, J., Li, F.F.: Siamese masked autoencoders. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 40676–40693. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/7ffb9f1b57628932518505b532301603-Paper-Conference.pdf
40. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
41. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
42. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)

43. He, R., Cascante-Bonilla, P., Yang, Z., Berg, A.C., Ordonez, V.: Improved visual grounding through self-consistent explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13095–13105 (2024)
44. He, R., Cascante-Bonilla, P., Yang, Z., Berg, A.C., Ordonez, V.: Learning from models and data for visual grounding (2024), https://arxiv.org/abs/2403.13804
45. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 646–661. Springer (2016)
46. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. arXiv preprint arXiv:2207.13532 (2022)
47. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The Kinetics Human Action Video Dataset (2017), https://arxiv.org/abs/1705.06950
48. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
49. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009), https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
50. Leclerc, G., Ilyas, A., Engstrom, L., Park, S.M., Salman, H., Madry, A.: FFCV: Accelerating training by removing data bottlenecks. In: Computer Vision and Pattern Recognition (CVPR) (2023), https://github.com/libffcv/ffcv/. commit 45f1274
51. Lehner, J., Alkin, B., Fürst, A., Rumetshofer, E., Miklautz, L., Hochreiter, S.: Contrastive tuning: A little help to make masked autoencoders forget. arXiv preprint arXiv:2304.10520 (2023)
52. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. arXiv preprint arXiv:2211.09552 (2022)
53. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. arXiv preprint arXiv:2303.16058 (2023)
54. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European conference on computer vision. pp. 280–296. Springer (2022)
55. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
56. Likhosherstov, V., Arnab, A., Choromanski, K., Lucic, M., Tay, Y., Weller, A., Dehghani, M.: Polyvit: Co-training vision transformers on images, videos and audio. arXiv preprint arXiv:2111.12993 (2021)
57. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)

58. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
59. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2016)
60. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
61. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. In: International Conference on Learning Representations (2016)
62. Lu, C.Z., Jin, X., Huang, Z., Hou, Q., Cheng, M.M., Feng, J.: Cmae-v: Contrastive masked autoencoders for video action recognition. arXiv preprint arXiv:2301.06018 (2023)
63. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
64. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: 4th International Conference on Learning Representations, ICLR 2016 (2016)
65. Mishra, S., Robinson, J., Chang, H., Jacobs, D., Sarna, A., Maschinot, A., Krishnan, D.: A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. arXiv preprint arXiv:2210.16870 (2022)
66. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., Oliva, A.: Moments in Time Dataset: One Million Videos for Event Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 502–508 (2020)
67. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
68. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
69. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
70. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
71. Parthasarathy, N., Eslami, S., Carreira, J., Hénaff, O.J.: Self-supervised video pretraining yields strong image representations. arXiv preprint arXiv:2210.06433 (2022)
72. Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B.: Learning features by watching objects move. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2701–2710 (2017)
73. Piergiovanni, A., Kuo, W., Angelova, A.: Rethinking video vits: Sparse video tubes for joint image and video learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2214–2224 (2023)
74. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)

75. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
76. Shrivastava, A., Selvaraju, R.R., Naik, N., Ordonez, V.: Clip-lite: Information efficient visual representation learning with language supervision. In: International Conference on Artificial Intelligence and Statistics. pp. 8433–8447. PMLR (2023)
77. Singh, N., Wu, C.W., Orife, I., Kalayeh, M.: Looking similar sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26907–26918 (2024)
78. Sriram, A., Gaidon, A., Wu, J., Niebles, J.C., Fei-Fei, L., Adeli, E.: Home: Homography-equivariant video representation learning. arXiv preprint arXiv:2306.01623 (2023)
79. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852. PMLR (2015)
80. Srivastava, S., Sharma, G.: Omnivec: Learning robust representations with cross modal sharing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1236–1248 (2024)
81. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
82. Tang, Y., Shimada, D., Bi, J., Xu, C.: Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue (2024), `https://arxiv.org/abs/2403.16276`
83. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Neural Information Processing Systems (NeurIPS) (2022)
84. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: European Conference on Computer Vision. pp. 516–533. Springer (2022)
85. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 98–106 (2016)
86. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 835–851. Springer (2016)
87. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14549–14560 (2023)
88. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14733–14743 (2022)
89. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.G.: Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6312–6322 (2023)
90. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: International Conference on Computer Vision (ICCV) (2015)

91. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
92. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
93. Wu, H., Wang, X.: Contrastive learning of image representations with cross-video cycle-consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10149–10159 (2021)
94. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
95. Xu, J., Wang, X.: Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10075–10085 (2021)
96. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3333–3343 (2022)
97. Zhang, B., Yu, J., Fifty, C., Han, W., Dai, A.M., Pang, R., Sha, F.: Co-training transformer with videos and images improves action recognition. arXiv preprint arXiv:2112.07175 (2021)
98. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
99. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
100. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)

A Implementation Details

ViC-MAE *architecture.* We will release code and model checkpoints, along with the specific training configurations. We followed previous training configurations that also worked well for our models [30, 40]. The pseudocode of ViC-MAE is also provided in Algorithm 1.

Algorithm 1: ViC-MAE PyTorch pseudocode.

```
# V[N, T, C, H, W] - minibatch (T=1 for images)
# tau: temperature coefficient
# clambda: contrastive coefficient

for V in loader:
    # Distant sampling
    f_i, f_j = random_sampling(V)
    # Patch embeddings and position encodings
    x_i = patch_embedd(f)
    x_i += pos_embedd
    # Mask out patches
    x_i, mask_i, ids_restore_i = random_masking(x)
    # Patchify, add pos_embed and mask out f_i ...
    # Forward frames on masked input
    x_i = frame_encoder(x_i) # [N, L_msk, D]
    x_j = frame_encoder(x_j) # [N, L_msk, D]
    # Pool features
    x_pool_i = pooling(x_i) # [N, D]
    x_pool_j = pooling(x_j) # [N, D]
    # Project and normalize
    p_i = l2_normalize(projector(x_pool_i), dim=1)
    z_j = l2_normalize(projector(x_pool_j), dim=1)
    # Predict pixels
    pred = decoder(x_i) # after adding mask tokens

    # compute pixel loss
    target = patchify(f_i)
    loss_pixel = (pred - target) ** 2
    loss_pixel = loss_pixel.mean(dim=-1) # [N, L]
    loss_pixel = (loss * mask).sum() / mask.sum()
    # compute contrastive loss
    loss_cons = ctr(p_i, z_j) + ctr(z_j, p_i)
    # compute final loss
    loss = loss_pixel + clambda * loss_cons


def ctr(p, z):
    # similarity matrix [N, N]
    sim = einsum('nl,nl->nn', p, z) * exp(tau)
    # compute info-nce loss
    labels = range(N) # positives are in diagonal
    loss = cross_entropy_loss(sim, labels)
    return 2 * loss
```

We follow the standard ViT architecture [27], which has a stack of Transformer blocks, each of which consists of a multi-head attention block and an Multi-Layer Perceptron (MLP) block, with Layer Normalization (LN). A linear projection layer is used after the encoder to match the width of the decoder. We use sine-cosine position embeddings for both the encoder and decoder. For the projection and target networks, we do average pooling on the encoder features and follow the architecture of Bardes *et al.* [6], which consists of a linear layer projecting the features up to twice the size of the encoder and two blocks of linear layers that preserve the size of the features, followed by batch normalization and a ReLU non-linearity.

We extract features from the encoder output for fine-tuning and linear probing. We use the class token from the original ViT architecture, but notice that similar results are obtained without it (using average pooling).

Video Loading. In order to prevent video loading from being a bottleneck on performance due to time spent on video decoding, we leverage the `ffcv` library [50], which we modify to support videos as a list of images in the WebP format. This allows us to significantly surpass the default PyTorch data loaders which can only read data in a synchronous fashion, resulting in the process being blocked until video decoding is complete. The use of `ffcv` allows to perform training without the need of sample repetition as done in OmniMAE [33] and ST-MAE [30] at the

cost of a significantly larger storage requirement. We will also release the code for `ffcv` to support videos.

Pre-training. The default settings can be found in Table 5. We do not perform any color augmentation, path dropping or gradient clipping. We initialize our transformer layer using xavier_uniform [35], as it is standard for Transformer architectures. We use the linear learning rate (lr) scaling rule so that lr = *base_* lr×batchsize / 256 [37].

End-to-end finetuning. We follow common practice for end-to-end finetuning. Default settings can be found in Table 6. Similar to previous work, we use layer-wise lr decay [40].

Linear evaluation. We follow previous work for linear evaluation results [30, 40]. As previous work has found we do not use common regularization techniques such as mixup, cutmix, and drop path, and likewise, we set the weight decay to zero [20]. We add an extra batch normalization layer without the affine transformation after the encoder features. Default settings can be found in Table 7.

Table 5: Pre-training setting.

config	value
optimizer	AdamW [60]
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.95 [15]
batch size	4096
learning rate schedule	cosine decay [59]
warmup epochs [37]	40
epochs	800
augmentation	hflip, crop [0.5, 1]
contrastive loss weight λ	0.025
contrastive loss schedule	0 until epoch 200 then 0.025

B Combining MAE with Negative-Free Methods.

We tried to combine MAE with instance discrimination learning methods by using the [CLS] token of the transformer as a global video feature representation. This representation allows us to use any instance discrimination learning loss without modifications to the underlying ViT transformer encoder. This combination works as follows: Sample two images from a video or an image and its transformed version I_i, I_j and perform patch-level masking. The two inputs are processed by the ViT model f_θ producing token representations $f_\theta(I_i) = \{x_i^{\text{CLS}}, x_i^1, x_i^2, \cdots, x_i^L\}$, where L is the sequence length of the transformer model. This is divided into two disjoint sets. The set $\{x_i^1, x_i^2, \cdots, x_i^L\}$ represents the local features of the input i

Table 6: End-to-end fine-tuning setting.

config	ViT/B	ViT/L
optimizer	AdamW	
optimizer momentum	β_1, β_2=0.9, 0.999	
base learning rate		
IN1K	3e-3	0.5e-3
P65	2e-3	
K400	1.6e-3	4.8e-3
SSv2	1e-3	
weight decay	0.05	
learning rate schedule	cosine decay	
warmup epochs	5	
layer-wise lr decay [4, 22]	0.65	0.75
batch size	1024	768
training epochs		
IN1K	100	50
P65	60	50
K400	150	100
SSv2	40	
augmentation	RandAug (9, 0.5) [23]	
label smoothing [81]	0.1	
mixup [98]	0.8	
drop path [45]	0.1	0.2

and are used for masked image modeling following Eq. 1. Then, the x_i^{CLS} token can be used as a global representation with a contrastive loss.

We experiment with this approach using the SimSiam loss [19] and the VicReg loss [6]. We review here these methods and how to combine them with MAEs, but the reader is referred to the original works for a more in-depth explanation of these methods [6, 19].

SimSiam. A combination of SimSiam and MAE, which we refer to as *MAE + SimSiam* uses the x_i^{CLS} token which represents the global video representation as follows: We pass x_i^{CLS} to a projector network $\mathcal{P}$ to obtain $p_i \triangleq \mathcal{P}(x_i^{\text{CLS}})/\|\mathcal{P}(x_i^{\text{CLS}})\|_2$. A similar procedure is followed for input j, but the global representation is not passed to the projector network $\mathcal{P}$ in order to obtain $z_j \triangleq x_i^{\text{CLS}}/\|x_i^{\text{CLS}}\|_2$. The SimSiam objective is then applied as follows:

$$\mathcal{L}_{p_i,z_j}^{\text{SimSiam}} = \|p_i - z_j\|_2^2 = 2(1 - p_i \cdot z_j). \tag{4}$$

VicReg. A combination of VicReg and MAE, which we refer to as *MAE + VicReg* uses the x_i^{CLS} token which represents the global video representation as follows: We pass it to a projector network $\mathcal{P}$ to obtain $p_i \triangleq \mathcal{P}(x_i^{\text{CLS}})/\|\mathcal{P}(x_i^{\text{CLS}})\|_2$, we repeat this procedure for input j using the target network $\mathcal{T}$ to obtain $z_i \triangleq \mathcal{T}(x_i^{\text{CLS}})/\|\mathcal{T}(x_i^{\text{CLS}})\|_2$. The loss is calculated at the embedding level on p_i

Table 7: Linear evaluation setting.

config	value
optimizer	SGD
base learning rate	0.1
weight decay	0
optimizer momentum	0.9
batch size	4096
learning rate schedule	cosine decay
warmup epochs	10
training epochs	90
augmentation	RandomResizedCrop

Table 8: Combining MAE and contrastive methods is not trivial. Linear evaluation on the ImageNet-1K dataset using types of contrastive learning. We use the [CLS] token as the global video representation and apply common contrastive methods, but these do not result on the best performance, which is obtained with our method.

Method	**ImageNet-1K**	
	Top-1	Top-5
MAE [40] + SiamSiam [19]	58.58	82.88
MAE [40] + VicReg [6]	63.86	84.07
ViC-MAE (ours)	**67.66**	**86.22**

and z_j. The inputs are processed in batches, let us denote $P = [p^1, \cdots, p^n]$ and $Z = [z^1, \cdots, z^n]$, where each p^m and z^m are the global representation of video m after the projector network and target network respectively in a batch of size n vectors of dimension d. Let us denote by p_l the vector composed of each value at dimension l in all vectors in P. The variance loss of VicReg is then calculated as follows:

$$v(P) = \frac{1}{d}\sum_{l=1}^{d} \max(0, \gamma - S(p_i, \epsilon)), \tag{5}$$

where $S(z, \epsilon) = \sqrt{\mathrm{Var}(z) + \epsilon}$ and γ is a constant target value for the standard deviation, fixed to 1. The covariance loss of VicReg can be calculated as:

$$c(P) = \frac{1}{d}\sum_{l \neq k}^{d} [\mathrm{Cov}(p^m)]_{l,k}^2, \tag{6}$$

where $\mathrm{Cov}(p^m) = \frac{1}{N-1}\sum_m (p^m - \bar{p})(p^m - \bar{p})^T$. The final VicReg loss over the batch is defined as:

$$\mathcal{L}_{p_i, z_j}^{\mathrm{VicReg}} = \frac{\lambda}{n}\|p_i - z_j\|_2^2 + \mu\left[v(P) + v(Z)\right] + \nu\left[c(P) + c(Z)\right]. \tag{7}$$

Table 9: Semi-supervised evaluation on ImageNet. We performed end-to-end finetuning using the settings in 6, but disable RandAug and MixUp for this experiment.

Percentage of data	**5%**	**10%**	**25%**	**50%**	**75%**	**100%**
MAE [40] + SimSiam [19]	7.15	23.41	39.73	54.94	62.88	67.44
MAE [40] + VicReg [6]	47.48	56.63	66.62	73.00	75.29	77.41
ViC-MAE (ours)	**50.25**	**58.22**	**67.65**	**73.97**	**75.80**	**77.89**

We perform experiments using these two combinations of MAE and contrastive losses as baseline comparisons for our method but found them to be underperforming with only contrastive or only masked methods. In other words, it is not trivial to adapt constrastive learning methods to be used in combination with masked autoencoders. See Table 8 for more details. For the contrastive learning part we experiment with two alternatives.

- *MAE + {SimSiam or VicReg}*. The predictor consists of the backbone network f_θ and a projector followed by a predictor as in Bardes *et al*. [6]. The target encoder consists of the backbone f_θ and the projector, which are shared between the two encoders.
- ViC-MAE. The predictor and the target networks share the same architecture consisting of the backbone network f_θ and a projector following Bardes *et al*. [6].

When using the MAE + {SimSiam or VicReg} combinations, we use the [CLS] token from the ViT architecture which is typically used to capture a global feature from the transformer network and is used to fine-tune the network for downstream tasks such as classification.

Combining MAE with negative-free representation learning is non trivial, and we set to test these by comparing our model with MAE models with alternative negative-free learning objectives Siamsiam [19] and VicReg [6]. We present our results using linear evaluation on Table 8. We use the [CLS] token as the global video representation for contrastive pre-training for 400 epochs. We can notice that competing methods underperform compared to our model which uses pooling of the local features by an absolute margin of $> 3\%$ over the *MAE + VicReg* model.

Semi-supervised evaluation on ImageNet. We also test ViC-MAE against negative-free representation learning methods on the problem of Semi-Supervised evaluation on the ImageNet dataset. The setting consists on training on a subset of the training data and testing on the whole validation data. We chose subsets of size 5%, 10%, 25%, 50%, 75% and 100% of the whole training set of ImageNet. We compare our model against MAE [40] + SimSiam [19], and MAE [40] + VicReg [6]. Results are shown on Table 9, and show the supperiority of ViC-MAE over simple combinations of contrastive learning and masked image modeling.

C Extra Ablations

Temporal representation learning. To investigate the temporal representation learned by our model, we conducted an ablation study on finetuned video models using the K400 dataset, examining performance through standard evaluation, frame shuffling, and frame repetition strategies. We evaluated and averaged 16 random permutations and exhaustive single-frame repetitions. Our findings reveal a decrease in accuracy from the original 87.8% to 78.8% with shuffled frames and 60% with repeated frames, underscoring our model's ability to implicitly learn temporal representations despite not being explicitly designed for temporal modeling. These results demonstrate the model's effectiveness in finetuning for temporal representation learning, highlighting its capacity to capture diverse temporal features.