V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning

Lorenzo Mur-Labadia[1], Matthew Muckley[2], Amir Bar[2], Mido Assran[2], Koustuv Sinha[2], Mike Rabbat[2], Yann LeCun[2], and Nicolas Ballas[2]

[1] Universidad de Zaragoza lmur@unizar.es
[2] FAIR at Meta

Abstract. We present **V-JEPA 2.1**, a family of self-supervised models that learns *dense*, high-quality, and temporally consistent representations for visual scenes in both images and videos. V-JEPA 2.1 combines four key ingredients: (i) a **Dense Predictive Loss**, a masking-based objective in which *all* tokens—visible context and masked tokens alike—contribute to the training loss, encouraging explicit spatial and temporal grounding; (ii) **Deep Self-Supervision**, which applies the self-supervised objective hierarchically at multiple intermediate encoder layers to improve representation quality; (iii) **Multi-Modal Tokenizers** that support unified training over images and videos; and (iv) effective **model and data scaling**. Empirically, V-JEPA 2.1 achieves state-of-the-art results on a range of benchmarks: 7.71 mAP on Ego4D for short-term object-interaction anticipation, 40.8 Recall@5 on EPIC-Kitchens for high-level action anticipation, and a 20% improvement in real-robot grasping success rate over VJEPA-2 AC. The model also achieves state-of-art performance in robotic navigation (5.687 ATE on Tartan Drive), depth estimation (0.307 RMSE on NYUv2 with a linear probe), and global recognition (77.7% on Something-Something-V2).

Keywords: Self-Supervised Learning · Dense Features · Video Learning

1 Introduction

World models hold the promise of enabling agents to perceive, predict, and plan effectively in the physical world [2, 34, 63]. At the core of these models lies the state-estimation problem: learning representations that reliably summarize the current world state from noisy perceptual inputs. Self-Supervised Learning (SSL) from video has recently emerged as a powerful route to this goal [3, 15], because it can exploit large-scale, label-free data to learn representations that capture scene geometry, dynamics, and intrinsic physical properties [28, 61].

Despite rapid progress, learning representations that simultaneously preserve *dense* spatio-temporal structure (needed for localization, geometry, and tracking) while also capturing *dynamics* and supporting *global* understanding (needed for high-level recognition) remains an open challenge. Among recent advances,

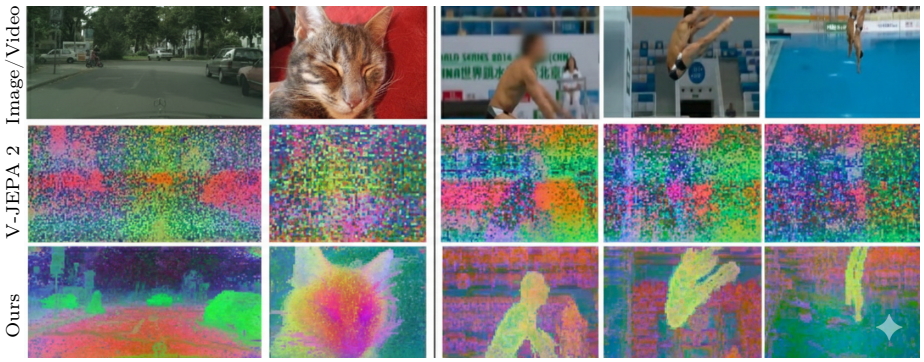


Fig. 1: V-JEPA 2.1 unlocks high-quality dense features. We compute PCA on patch features from V-JEPA 2 (ViT-g) and V-JEPA 2.1 (ViT-G). V-JEPA 2.1 learns *spatially structured, semantically coherent, and temporally consistent features.*

Joint Embedding Predictive Architectures (JEPA) [43]—and in particular the V-JEPA family [3, 8]—have demonstrated strong global video understanding, especially in scenarios requiring modeling of motion and temporal dynamics, showing promise for enabling prediction and planning in embodied agents [3]. However, as illustrated in Figure 1, previously learned representations fail to capture fine-grained local spatial structure. In contrast, other SSL approaches such as DINO [15, 51, 61] yield high-quality dense features, but are primarily image-based and therefore do not directly learn temporal dynamics from video.

In this work, we study self-supervised learning with a latent mask-denoising objective, where the model predicts masked segments of an image or video directly in a learned representation space. Our central finding is that high-quality *dense* spatio-temporal features—preserving fine-grained spatial layout and motion dynamics—do not emerge reliably when the prediction loss is applied only to masked regions. Instead, extending the predictive loss to the *entire* input, both masked and unmasked segments, substantially improves dense representations.

Building on this insight, we introduce **V-JEPA 2.1**, a self-supervised approach for learning unified image and video representations. V-JEPA 2.1 uses a *dense predictive loss* applied to *all tokens* (both visible context and masked tokens), preventing visible tokens from acting as global aggregators—an effect that is key to improving dense feature quality (Figure 3). Additionally, we find that *deep self-supervision*—applying the loss hierarchically at multiple intermediate encoder layers to provide training signals throughout the network—yields consistent gains on both dense and global downstream tasks. To enable native joint training across modalities, we use *modality-specific learned tokenizers* for images and videos within a single shared encoder. Finally, we show these improvements *scale with data and model capacity*: expanding the image component from 1M to 142M images using VisionMix-163M and scaling the model from 300M to 2B parameters leads to systematic downstream gains.

Empirically, V-JEPA 2.1 achieves state-of-the-art performance on *predictive* video benchmarks spanning both fine-grained and semantic forecasting: it reaches 7.71 mAP on Ego4D short-term object-interaction anticipation, which requires predicting *where* and *when* interactions will occur (localized interaction regions and time-to-interaction), and 40.8 Recall@5 on EPIC-Kitchens-100 action anticipation, which evaluates the ability to forecast upcoming actions from partial temporal context. Better dense features also improve performance on world modelling tasks: V-JEPA 2.1 leads to a +20% success rate compared to VJEPA-2 AC [3] on grasping when deploying our model on a real Franka arm in a new environment zero-shot; and on robot navigation, V-JEPA 2.1 achieves state-of-the-art performance (5.687 ATE on Tartan Drive) with 10x faster planning speed compared to previous work [7]. Beyond prediction and planning, V-JEPA 2.1 also delivers strong performance in both *dense* and *global* understanding tasks. For dense tasks, V-JEPA 2.1 ViT-G sets a new state-of-the-art in linear-probe monocular depth estimation (0.307 RMSE on NYUv2), achieves competitive linear-probe semantic segmentation (85.0 mIoU on Pascal VOC), and produces temporally consistent features for video object segmentation (72.7 $\mathcal{J}\&\mathcal{F}$-Mean on YouTube-VOS). At the global level, it also attains state-of-the-art accuracy in action recognition (77.7% on SSv2). We release V-JEPA 2.1 models (ViT-g/G, 1B/2B) and two distilled variants (ViT-B/L, 80M/300M).

2 Related work

Self-Supervised Learning. Early SSL in vision relied on hand-crafted pretext tasks, such as predicting patch positions [22, 50], inpainting [52], or rotation [30]. The field shifted toward joint-embedding architectures to learn transformation invariance via contrastive [18, 35], non-contrastive [9, 33], or clustering-based [14] objectives. The adoption of Vision Transformers (ViTs) [23] led to Masked Image Modeling (MIM), which operates in either pixel space [72] or the transformer's latent space [6]. Modern state-of-the-art models often combine view-invariance with MIM objectives [51, 79]. Recently, JEPA [43] has formalized representation learning as predicting missing latent information, demonstrating high efficiency across images [2], audio [4], and video [8].

Video Models. Early video research utilized motion cues for pretext tasks like ego-motion prediction [40] or temporal patch consistency [69]. While large-scale datasets like Kinetics [17] initially favored supervised 3D ConvNets [65], efficiency concerns led to (2+1)D convolutions [66] and Video Transformers [1, 11]. Modern Video SSL has evolved from temporal order verification [46] to masked modeling extensions like VideoMAE [27, 64]. Scaling these approaches has produced generalist video encoders [16, 68], with JEPAs proving particularly effective for large-scale world modeling [3]. Additionally, incorporating weak language supervision has enabled these models to tackle language-video tasks [70, 77].

Learning Dense Features. Dense representation learning often utilizes local objectives, such as spatio-temporal consistency [39], spatial alignment be-

tween crops [10], or region-based contrastive learning [38, 71]. Recent distillation methods like AM-RADIO [57] and DINOv3 [61] use cosine similarity and Gram matrix regularization to further refine these features. Architecturally, the Dense Prediction Transformer (DPT) [56] addressed the token-to-pixel challenge through a "reassemble" operation that converts ViT sequences into multi-scale feature maps. This framework, alongside hierarchical backbones like Swin [45] and MViT [26], has become the standard for dense downstream tasks, serving as the foundation for modern models like Depth Anything [74].

3 Methodology

3.1 Preliminaries: Joint-Embedding Predictive Architectures

Joint-Embedding Predictive Architecture (JEPA) [43] is a self-supervised learning framework designed to learn representations of data by making predictions in a learned latent space, rather than directly in the observation (input) space. Both corrupted and clean versions of the same input are processed by the encoder $E_\theta(\cdot)$, and the predictor $P_\phi(\cdot)$ then trained to predict the representation of the clean input from the representation of the corrupted input. Given that the encoder and predictor are learned simultaneously, the system admits a trivial solution in which the encoder $E_\theta(\cdot)$ outputs a constant vector regardless of its input. To avoid this representation collapse, explicit [5, 9, 48] or implicit [2, 32] regularization is used to promote representations preserving input information.

In this work, we build upon the V-JEPA family of models, where video representations are learned through a mask-denoising objective in the representation space [3, 8]. The V-JEPA objective aims to predict the representation of a video y from another view x of that video that has been corrupted through masking, i.e., from which patches have been randomly dropped. The encoder $E_\theta(\cdot)$ processes the masked video x, and outputs an embedding vector, or context tokens, for each visible patch. The outputs of the encoder are concatenated with a set of learnable mask tokens Δ_y that specify the spatio-temporal position of the masked patches. The predictor network processes the combined tokens sequence, and outputs an embedding vector for each input token. The encoder and predictor are trained by minimizing the following objective:

$$\mathcal{L}_{\text{predict}} = \frac{1}{|M|} \sum_{i \in M} \| P_\phi(E_\theta(x), \Delta_y)_i - \text{sg}(E_{\overline{\theta}}(y)_i) \|_1, \tag{1}$$

where M is a set containing the masked patch indexes from the x view. The loss uses a stop-gradient operator, sg, to prevent representation collapse [33] and an exponential moving average $\overline{\theta}$ of θ is used to update the weight of the encoder $E_{\overline{\theta}}(y)$ processing the video y. $\mathcal{L}_{\text{predict}}$ is only applied on masked tokens, and not on the context tokens. Both encoder $E_\theta(\cdot)$ and predictor $P_\phi(\cdot)$ are parametrized with Vision Transformer [23] and we rely on the same masking strategy as [3].

While V-JEPA has proven to be an effective approach for understanding global semantic information from video, predicting future actions, and planning

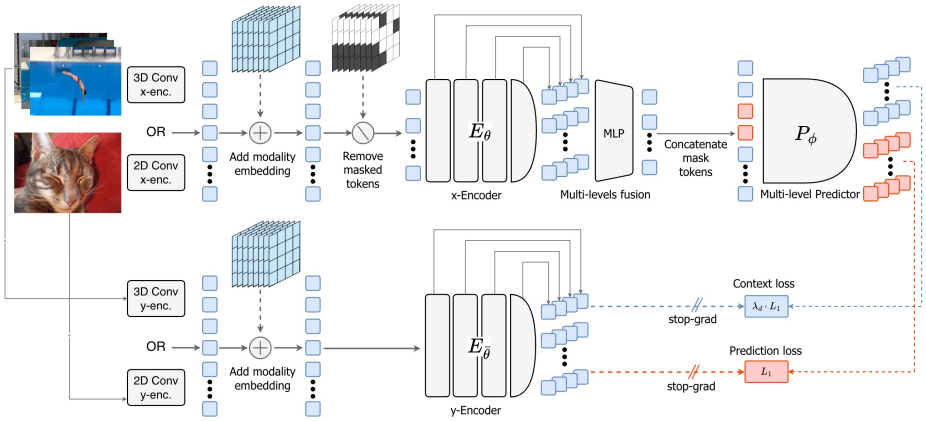


Fig. 2: V-JEPA 2.1 Architecture. Images and videos are processed by respectively either a 2D or 3D Convolutional patch embedding. Then, 3D Rotational Positional Encoding (RoPE) and learnable modality embedding are added. The $x-$encoder processes the visible tokens and outputs multi-level embeddings formed by concatenating the normalized output from intermediate encoder blocks. Then, a MLP fuses this multi-level representation and reduces its dimensionality. These context tokens are concatenated with learnable mask tokens carrying spatio-temporal positional information. The predictor processes the combined sequence and produces multi-level predictions for the masked tokens. Training uses two different losses: (i) an L1 loss on masked-token predictions (the original V-JEPA objective), and (ii) a distance-weighted L1 loss on nearby context tokens, both supervised using the y-encoder multi-level outputs.

to reach specific goals [3, 8], the previous V-JEPA 2 model obtained very low performance on dense tasks when using a simple linear probing protocol, such as semantic segmentation (22.2 mIoU on ADE20K) or depth estimation (0.682 RMSE on NYUv2). Qualitatively, these feature maps are noisy and show only fragmented local spatial structure, as shown in Figure 1 and 3.

3.2 V-JEPA 2.1: Improving Dense Video SSL Features

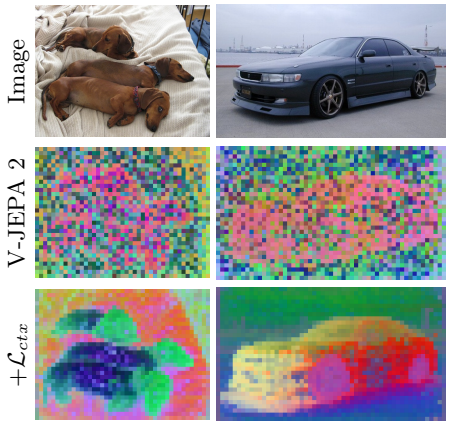
To this end, we introduce V-JEPA 2.1, a self-supervised training recipe for learning representations that combine high-quality dense local features with global semantic understanding. Our key algorithmic contributions are (1) Dense Prediction loss that applies self-supervision on both masked and unmasked tokens and, (2) Deep Self-Supervision of the encoder intermediate layers via a multi-level predictor. Additionally, we explore (3) a Multi-Modal Tokenizer with modality-specific patch embeddings for images and videos; (4) Data Scaling through a more diverse and balanced image–video training distribution; and (5) Model Scaling to ViT-G, enabling state-of-the-art downstream performance and effective (6) distillation to smaller models (ViT-L, ViT-B).

VJEPA 2.1 Architecture. We illustrate the V-JEPA 2.1 architecture in Figure 2. The input, either an image or a video, is projected into a sequence of

Table 1: Impact of individual components of V-JEPA 2.1 training recipe. Results on image classification (IN1K), video classification (SSv2), depth estimation (NYU), and semantic segmentation (ADE20K). Introducing the Context Loss improves dense tasks but reduces classification performance on SSv2. Incorporating our Deep Self-Supervision restores the classification performance. The VisionMix 163M dataset, the Multi-Modal Tokenizer, and scaling model size further improve performance.

IN1K Acc.	SSv2 Acc.	NYU RMSE ↓	ADE20K mIoU	
82.2	72.8	0.682	22.2	V-JEPA 2
72.6	62.5	0.474	33.8	+ Context Loss
80.8	72.1	0.463	38.6	+ Multi-level Pred.
81.6	72.6	0.418	40.8	+ Vision Mix
81.6	72.6	0.415	41.4	+ Multi-modal Tok.
84.8	76.1	0.365	47.1	+ Model Scaling
85.5	**77.7**	**0.307**	**47.9**	+ Cool-down.

Fig. 3: Influence of the context loss $\mathcal{L}_{ctx}$. We show PCA visualizations of the feature map representations learned with V-JEPA 2 and with a model trained using V-JEPA 2 plus our $\mathcal{L}_{ctx}$ loss.



embedding vectors, or tokens, using a modality-specific patch embedding. Mask corruption is then applied to the sequence by randomly dropping patch tokens. The x-encoder processes the remaining visible context tokens and outputs representations from multiple encoder levels in addition to the final output. The multi-level representations are then concatenated along the channel axis and fed to an MLP to reduce their dimensionality. Context tokens are concatenated, along the sequence axis, with learnable mask tokens that carry spatio-temporal positional information of the masked patches. The predictor processes the combined sequence and produces multi-level predictions for each token. Training uses two different losses: (i) an L1 loss on masked-token predictions (the original V-JEPA objective), and (ii) a distance-weighted L1 loss for context tokens. Both use the y-encoder outputs as targets, which process the unmasked sequence of patches from the input images or videos. Losses are applied to several intermediate representation levels in addition to the encoder output.

Next, we describe in detail each component and analyze its impact on downstream performance. As we show in Table 1, we ablate V-JEPA 2.1 performance on dense prediction tasks (ADE20K, NYUv2) with linear probing [61] and global recognition tasks (SSv2, ImageNet) with attentive probing [3].

1. Dense Prediction Loss. We attribute the lack of local structure in the feature maps to the absence of supervision on patches that are not masked, i.e. the context patches. In the previous V-JEPA 2, the predictor $P_\phi(\cdot)$ processed both context and masked tokens, yet the loss was applied only to masked tokens (Eq. 1). As a result, the model was not encouraged to preserve local information in context tokens and instead prioritized global aggregation to minimize

$\mathcal{L}_{\text{prediction}}$, similarly to register tokens [21]. To address this limitation, we introduce a novel context loss $\mathcal{L}_{\text{ctx}}$ that enforces consistency between the predictor outputs for context tokens and the corresponding $y-$encoder representations:

$$\mathcal{L}_{\text{ctx}} = \frac{1}{|C|} \sum_{i \in C} \lambda_i \| P_\phi(E_\theta(x), \Delta_y)_i - \text{sg}(E_{\overline{\theta}}(y)_i) \|_1, \quad \lambda_i = \frac{\lambda}{\sqrt{\text{d}_{\min}(i, M)}} \tag{2}$$

where C is the set of indexed context tokens. Additionally, for a given patch i, the context loss $\mathcal{L}_{\text{ctx}}$ is weighted by the inverse square root of its minimum spatio-temporal distance $\text{d}_{\min}$ (in number of blocks) to any masked token M in the sequence. This weighting emphasizes patches near masked regions by enforcing local continuity between masked and context areas, yielding a good trade-off between segmentation and action recognition performance.

Figure 3 illustrates the impact of $\mathcal{L}_{\text{ctx}}$ on the learned representations. With the context loss, local structure now clearly appears in the feature maps, noisy artifacts are removed, and similar semantic parts (e.g., head of the dogs, wheel of the car) are mapped to the same PCA components. Introducing $\mathcal{L}_{ctx}$ with our weighted scheme improves the performance on dense vision tasks (22.2 $\rightarrow$ 33.9 mIoU on ADE20k, 0.682 $\rightarrow$ 0.473 RMSE on NYUv2). However, Table 1 shows that there is still a degradation in video understanding (72.8 $\rightarrow$ 62.5 % Acc. on SSv2) and image classification (82.2 $\rightarrow$ 72.6 % Acc. on IN1K).

2. Deep Self-Supervision. We self-supervise the encoder representation not only at the output but also at multiple intermediate levels. We first concatenate, along the channel dimension, the outputs of three intermediate x-encoder blocks, in addition to the output layer. Then, a lightweight MLP fuses these multi-level representations and reduces their dimensionality. The predictor processes the fused multi-level sequence of context and mask tokens and produces four outputs corresponding to the four encoder layers. Both the prediction loss and the context loss are then applied at each one of these four levels.

Deep Self-Supervision leads to significant improvement on downstream performance for both global and dense tasks, as shown in Table 1. Furthermore, it allows local information to flow towards the final layers, effectively removing the need for intermediate layers in dense downstream tasks as we show in Suppl. Material. Deep Self-Supervision allows recovering the global understanding capabilities of V-JEPA 2 (62.5$\rightarrow$72.1 % Acc. on SSv2, 72.6$\rightarrow$80.8 % Acc. on IN1K) while improving on dense tasks with the context loss (33.8$\rightarrow$38.6 mIoU on ADE20K, 0.474$\rightarrow$0.463 RMSE on NYU).

3. Scaling Image Data. DINOv2 [51] introduced a cluster-based retrieval strategy to select images from a large pool of raw internet data, resulting in a curated set of 142 million images. Using a similar approach, V-JEPA 2 [3] collected and curated video scenes from YT1B videos [75], combined with other publicly available video datasets, yielding a large-scale collection of 19 million video samples from the internet, corresponding to 1.6 million hours. Both works demonstrated the positive effect of data scaling for SSL pretraining.

Building on these insights, we construct our VisionMix163M Dataset, combining large-scale curated sources from these two prior works. Replacing the 1M-image ImageNet subset from V-JEPA 2 pretraining data with LVD-142M provided a broader and more diverse appearance data distribution. As we report in Table 1, the improvement of training on a more diverse and extended dataset is beneficial in all tasks (72.1 $\rightarrow$ 72.6 % Acc. on SSv2, 80.8 $\rightarrow$ 81.6 % Acc. on ImageNet, 38.6 $\rightarrow$ 40.8 mIoU on ADE20K, 0.463 $\rightarrow$ 0.418 RMSE on NYUv2).

4. Multi-Modal Tokenizer. Previous joint image–video training approaches, such as V-JEPA 2, were suboptimal because they only used a single 3D convolution for patch embedding, duplicating images temporally and treating them as 16-frame static videos. This increased computational cost and introduced an incorrect representational bias (i.e., images were interpreted as static videos). Instead, we introduce a multi-modal tokenizer, which applies a 3D convolution of $16 \times 16 \times 2$ for processing videos and a 2D convolution of 16×16 for images, allowing each modality to be processed in its native form.

We also add a modality-learnable token to both the encoder and the predictor inputs, which explicitly encodes whether the input comes from the image or video pathway. These learnable tokens condition the processing on the input modality, helping the model disentangle stronger static appearance cues in images and temporal motion information in videos. Quantitatively, both components have a positive effect on dense-task performance, improving mIoU on ADE20K 40.8 $\rightarrow$ 41.4, while performance on action or object recognition remains stable.

5. Scaling Model Size to ViT-G (2B) and High-Resolution Cool-Down. Next, we explore the effect of model scaling and high-resolution cooldown. Scaling the V-JEPA encoder from a ViT-L with 300 million parameters to a ViT-G with 2 billion parameters leads to significant improvements on all downstream tasks (72.6 $\rightarrow$ 76.1 % Acc. on SSv2, 81.6 $\rightarrow$ 84.8 % Acc. on ImageNet, 41.4 $\rightarrow$ 47.1 mIoU on ADE20K, 0.415 $\rightarrow$ 0.365 RMSE on NYUv2), as Table 1 shows.

Additionally, introducing a cool-down phase for the learning rate, during which we increase both the spatial resolution of images (from 256×256 to 512×512) and the spatio-temporal resolution of videos (from 16 frames at 256×256 to 64 frames at 384×384), further improves performance on all tasks: final Accuracy of 77.7 % on SSv2, 85.5 % on ImageNet and 47.9 mIoU on ADE20K. The benefits of a final cool-down with longer videos are particularly strong in depth estimation (0.365 $\rightarrow$ 0.307 RMSE on NYUv2).

6. Family of distilled models. We scaled model size to a large ViT-G (2B) not only for top-tier performance but also to enable distillation [36] into smaller models (ViT-L, 300M; ViT-B, 80M) appreciated by the community. The distillation protocol adapts the pretraining recipe by replacing the EMA target encoder with a frozen teacher, maintaining an EMA copy of the student as the final model, and computing the loss only on the teacher's last-layer representations without deep supervision. The predictor is reduced to 12 blocks with a final projection

Table 2: Short-Term Object Interaction Anticipation on Ego4D. Mean Average Precision (mAP) [31].

Model	N	N+V	N+δ	All
Literature				
StillFast [55]	20.2	10.3	7.17	3.96
STAformer [49]	29.4	15.3	9.94	5.67
Same protocol				
DINOv2 ViT-g [51]	30.6	16.8	9.37	5.44
DINOv3 ViT-7B [61]	**33.8**	**17.4**	10.3	5.68
V-JEPA 2 ViT-g [3]	27.2	14.9	10.4	6.02
V-JEPA 2.1 ViT-g	28.7	15.5	11.4	6.75
V-JEPA 2.1 ViT-G	30.9	17.2	**12.8**	**7.71**

Table 3: Action Anticipation on EK100. Mean-class Recall@5 for verb, noun and action on EK100 val.

Method	Param.	Verb	Noun	Action
Literature				
InAViT [59]	160M	51.9	52.0	25.8
Video-LLaMA [76]	7B	52.9	52.0	26.0
PlausiVL [47]	8B	55.6	54.2	27.6
Same protocol				
V-JEPA 2 ViT-g [3]	1B	63.6	57.1	39.7
V-JEPA 2.1 ViT-g	1B	63.6	56.2	38.4
V-JEPA 2.1 ViT-G	2B	**64.3**	**59.9**	**40.8**

layer, while all other training hyperparameters remain unchanged. More details on the training and distillation procedures are provided in the Suppl.Material.

4 Results

In this section, we evaluate the performance of V-JEPA 2.1 on a large variety of downstream tasks. Throughout the experiments, we employ V-JEPA 2.1 as a *frozen encoder*, demonstrating the versatility of its features.

4.1 Short-Term Object Interaction Anticipation (STA). STA predicts future object interaction in egocentric videos, including the next active object bounding box b, the object noun N, the action verb V and the time until contact (δ). This task evaluated fine-grained 2D localization, semantic understanding of objects, temporal reasoning, and the ability to forecast user intentions. Using the Ego-4D STA benchmark [31], V-JEPA 2.1 achieves absolute state-of-the-art performance with an overall mAP of 7.71 as reported by Table 2. This represents a relative improvement of approximately 35% over the previous best method [49], which introduces task-specific training components. This gain is primarily driven by improved understanding of the next action 17.2 MAP_{V} and the time to 12.8 mAP $_{\delta}$, showing the strong *predictive* capabilities of V-JEPA 2.1.

4.2 Action Anticipation. This task predicts future actions given a contextual video clip leading up to some time before the action. Using the EPIC-Kitchens-100 (EK100) benchmark [20], we show that V-JEPA 2.1 outperforms the action anticipation performance of V-JEPA 2, and sets a new state-of-the-art for the task. We employ the same protocol as V-JEPA 2 and use an attentive probe trained on top of the frozen V-JEPA 2.1 encoder and predictor to anticipate future actions. Table 3 summarizes the results on the validation set of the EK100 action anticipation benchmark. We compare V-JEPA 2.1 ViT-g 1B and ViT-G 2B encoders with V-JEPA 2 ViT-g 1B encoder. Both leverage 32 frames with

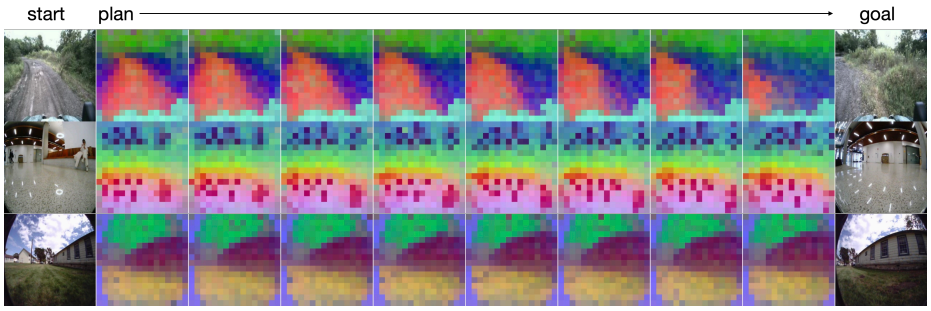


Fig. 4: Planning Navigation Trajectories in Latent Space. V-JEPA 2.1 enables 10× faster and more accurate navigation planning compared to [7]. We show PCA visualizations of 8 denoising steps of the planned latent trajectory.

Table 4: Open Loop Navigation Planning. We report average planning time, normalized Average Trajectory Error (ATE) and Relative Trajectory Error (RTE).

Model	Time	Recon		Tartan Drive		Scand		Sacson		Average	
		ATE	RTE	ATE	RTE	ATE	RTE	ATE	RTE	ATE	RTE
NWM [7]	103.2 (s)	1.146	0.339	5.831	1.219	1.113	0.297	4.037	0.928	3.032	0.696
V-JEPA 2.1 ViT-g	10.6 (s)	**1.146**	**0.338**	5.758	1.200	1.094	0.299	**3.904**	**0.921**	**2.975**	0.690
V-JEPA 2.1 ViT-G	**10.6** (s)	1.179	0.340	**5.687**	**1.187**	**1.038**	**0.285**	4.054	0.938	2.990	**0.688**

8 frames per second at resolution 384 × 384 as video context. V-JEPA 2.1 is comparable to V-JEPA 2 on the same 1B model size, but show scaling of the performance to the 2B models size, setting a new absolute state-of-the-art performance at 40.8 Action Recall@5, corresponding to a +2.8% improvement.

4.3 Navigation Planning. We adopt the navigation planning setup of NWM [7] and train a latent world model on top of the V-JEPA 2.1 representations. Specifically, given an agent's most recent observation and a goal location specified by an image, we predict a 2-second navigation trajectory toward the goal using a Conditional Diffusion Transformer (CDiT). Our initial experiments indicate that directly training a diffusion model in the high-dimensional embedding space of V-JEPA 2.1 is challenging: the representations are high-dimensional (at least 80× larger than those of an SD-VAE), injecting noise in arbitrary directions can easily push samples off the data manifold. To improve robustness, we introduce two modifications. First, we train the model to predict a clean representation rather than noise [44]. Second, we use DDIM sampling [62], which is less stochastic than DDPM [37] and empirically improves stability. We find that V-JEPA 2.1 enables *more accurate planning* while achieving 10× *faster planning speed* than the previously used SD-VAE [58] without sacrificing performance. Table 4 reports our results, with qualitative examples shown in Figure 4.

4.4 Robotic Arm Planning. We train a frame-causal action-conditioned predictor, following the same protocol of [3], where the predictor is a 24-layer trans-

Table 5: Planning Performance. Closed-loop robot manipulation of a cup using MPC. Results are averaged over 10 tasks per skill. The improved depth understanding from V-JEPA 2.1 dense features lead to a 20% improvement in Grasp success rate.

Method	#Samples	Iter.	Horizon	Time	**Reach**	**Grasp**	**Pick-&-Place**
VJEPA 2 [3]	800	10	1	3 sec.	100%	60%	80%
VJEPA 2.1	800	10	1	3 sec.	100%	70%	80%
	300	15	8	14 sec.	100%	**80%**	80%

Table 6: Performance on dense visual tasks. We report the RMSE for depth estimation (NYUv2, KITTI), mIoU for semantic segmentation (ADE20K, Cityscapes, VOC12), and the $\mathcal{J}\&\mathcal{F}$-Mean for video object segmentation (DAVIS, YouTube-VOS). V-JEPA 2.1 VIT G demonstrates strong performance on dense vision tasks, outperforming DINOv3 ViT-7B on NYUv2 depth estimation.

Method	**Param.**	**Depth**		**Segmentation**			**VOS**	
		NYUv2↓	KITTI↓	ADE20k	Citysc.	VOC	DAVIS-S	YT VOS-S
7B models								
Web-DINO [25]	7B/14	0.466	3.158	42.7	68.3	76.1	57.2	43.9
DINOv3 [61]	7B/16	**0.309**	**2.346**	**55.9**	**81.1**	**86.6**	**71.1**	**74.1**
Models less than 2B parameters								
PEcore [13]	2B/14	0.590	4.119	38.9	61.1	69.2	48.2	34.7
PEspatial [13]	2B/14	0.362	3.082	49.3	73.2	82.7	68.4	68.5
AM-RADIOv2.5 [57]	1B/14	0.340	2.918	53.0	78.4	85.4	66.5	70.1
InternVideo2-1B [70]	1B/14	0.471	4.739	28.4	35.6	65.2	50.6	51.2
SigLIP 2 [67]	1B/16	0.494	3.273	42.7	64.8	72.7	56.1	52.0
DINOv2 [51]	1B/14	0.372	2.624	49.5	75.6	83.1	63.9	65.6
DINOv3 ViT-H+ [61]	0.8B/16	0.352	2.635	**54.8**	**79.5**	**85.8**	**71.1**	**74.0**
V-JEPA2 ViT-g [3]	1B/16	0.642	4.650	24.4	45.9	63.9	52.5	53.7
V-JEPA 2.1 ViT-g	1B/16	0.350	2.601	47.8	71.8	84.7	68.1	72.3
V-JEPA 2.1 ViT-G	2B/16	**0.307**	**2.461**	47.9	73.5	85.0	69.0	72.7

former network trained on the Droid raw dataset [42] using both a teacher-forcing and two-step rollout loss. The model is then deployed in our lab, zero-shot, on a table-top Franka Panda robot arm with a parallel-jaw gripper, and evaluated on Reach, Grasp, and Pick-and-Place tasks with visual goal specification via model-predictive control.

Success rates are reported in Table 5. VJEPA 2.1 exhibits an improved depth understanding, leading to a 10% improvement in the success rate of Grasp compared to VJEPA 2 (see. Suppl.Material). Moreover, we find that the VJEPA 2.1 model unlocks the benefit of planning with slightly longer horizons, perhaps due to the improvement in dense features afforded by our model. In particular, by reducing the number of CEM samples and planning over 8 steps, we observe an additional 10% improvement in the success rate on Grasp.

4.5 Depth Estimation and Semantic Segmentation. We evaluate V-JEPA 2.1 on single-image semantic segmentation and depth estimation, two tasks that

Table 7: Performance on global tasks. We report Top-1 classification accuracy on action recognition benchmarks (SSv2, Diving-48, and K400) and image classification (IN1K). V-JEPA 2.1 VIT-G achieves state-of-the-art performance on the SSv2 action recognition tasks while being competitive on appearance understanding tasks.

Method	Param.	Motion Understanding		Appearance Understanding	
		SSv2	Diving-48	K400	IN1K
Results Reported in the Literature					
InternVideo2-1B [70]	1B	67.3	–	87.9	–
VideoPrism [77]	1B	68.5	71.3	87.6	–
DINOv3 ViT-H+ [61]	0.8B	69.8	–	86.7	87.9
DINOv3 [61]	7B	70.1	–	87.8	**88.4**
Vision Encoders Evaluated Using the Same Protocol					
DINOv2 [21]	1.1B	50.7	82.5	83.6	86.1
PE_{core} G [13]	1.9B	55.4	76.9	88.5	87.6*
SigLIP2 [67]	1.2B	49.9	75.3	87.3	88.0
$\text{InternVideo2}_{2\text{s}}$-1B [70]	1B	69.7	86.4	**89.4**	85.8
V-JEPA ViT-H [8]	600M	74.3	87.9	84.5	80.0
V-JEPA 2 ViT-g [3]	1B	77.3	**90.2**	87.3	85.1
V-JEPA 2.1 ViT-g	1B	76.9	89.0	87.0	84.8
V-JEPA 2.1 ViT-G	2B	**77.7**	89.2	87.7	85.5

require fine-grained understanding of the image spatial structure. Following the protocol of DINOv3 [61], we train a dense linear projection on top of the frozen final-layer features of the V-JEPA 2.1 encoder.

As shown in Table 6, V-JEPA 2.1 ViT-G achieves state-of-the-art linear-probe monocular depth estimation, reaching 0.307 RMSE on NYUv2 [60] and 2.461 RMSE on KITTI [29], the best among models under 2B parameters. On NYUv2, V-JEPA 2.1 surpasses prior image encoders, including DINOv3 ViT-7B (0.309 RMSE on NYU) and $\text{PE}_{\text{spatial}}$-G (0.362 RMSE on NYU), and significantly outperforms video encoders like InternVideo2-1B (0.471 RMSE) and V-JEPA 2 (0.642 RMSE). For semantic segmentation, V-JEPA 2.1 remains highly competitive, obtaining 85.0 mIoU on VOC12 [24], 73.5 mIoU on Cityscapes [19] and 47.9 mIoU on ADE20K [78]. Compared with V-JEPA 2, we observe strong gains across datasets (+23.4 ADE20K, +27.6 Cityscapes, and +20.7 VOC12); showing the benefits of context supervision and the multi-level predictor. Performance on ADE20K and Cityscapes remains slightly below that of the best image encoders, likely due to the scarcity of highly cluttered scenes in VisionMix, limiting exposure to the fine-grained segmentation complexity of these benchmarks.

4.6 Video Object Segmentation (VOS). We assess the temporal consistency of V-JEPA 2.1 representations with the VOS task. Following [61], we adopted a non-parametric label propagation approach [39] that matches local patch features across frames using cosine similarity in the embedding space. We evaluate on DAVIS 2017 [54] and YouTube-VOS [73] datasets, reporting the standard $\mathcal{J}\&\mathcal{F}$-mean metric [53] that jointly measures the region similarity and contour

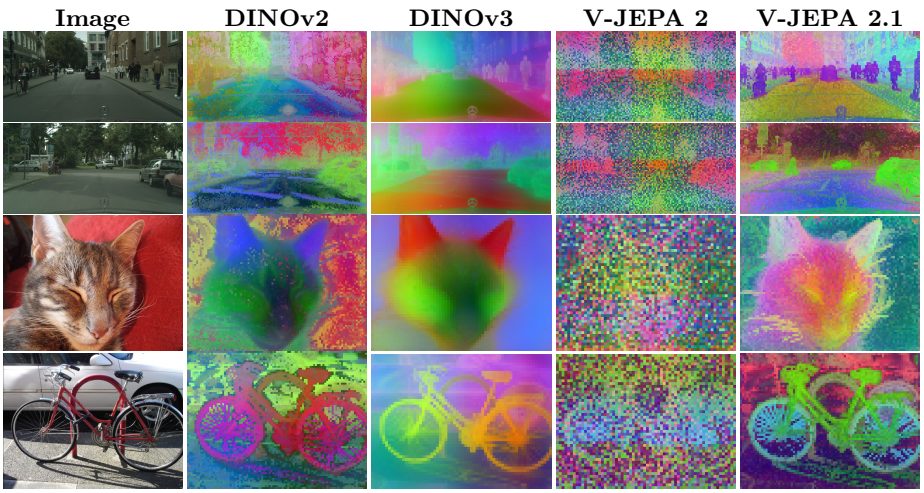


Fig. 5: Comparison of dense features produced by different SSL methods on a single image. We compare DINOv2 ViT-g [51], DINOv3 ViT-H+ [61] and V-JEPA 2 ViT-g [3] against V-JEPA 2.1 ViT-G. Images are resized to a shorter side of 1024 pixels and processed with the 2D convolutional image tokenizer.

accuracy. As we show in Table 6, V-JEPA 2.1 achieves 69.0 $\mathcal{J}\&\mathcal{F}$ on DAVIS-17 and 72.7 $\mathcal{J}\&\mathcal{F}$ on YouTube-VOS datasets, obtaining the second best performance and surpassing all prior encoders except DINOv3. These results highlight the temporal consistency of the V-JEPA 2.1 features, which enable stable object tracking despite significant appearance changes across frames.

4.7 Video and Image Classification. We assess the quality of V-JEPA 2.1 representations for high-level scene understanding on action recognition and object classification benchmarks. Table 7 reports our results using a 4-layer attentive probe trained on frozen encoder features, comparing V-JEPA 2.1 against a diverse set of visual encoders. V-JEPA 2.1 ViT-G achieves a top-1 Accuracy of 77.7%on SSv2, setting a new absolute state-of-the-art on the task compared to 77.5% for InternVideo2 full fine-tuning, 77.3% for V-JEPA 2, 70.1% for DINOv3 ViT-7B and 69.7% for InternVideo2 using the same protocol, demonstrating a strong understanding of video dynamics. Our model is also highly-competitive in appearance-based tasks, reaching 87.7% on K400 and 85.5% on IN1K.

4.8 Qualitative results of V-JEPA 2.1 Dense Features. Figure 5 compares dense feature maps extracted from a single image using V-JEPA 2.1 ViT-G, DINOv2 ViT-g [51], DINOv3 ViT-H+ [61], and V-JEPA 2 ViT-g [3]. Each feature space is projected to three dimensions via PCA. This highlights how V-JEPA 2.1 learns dense representations that are spatially structured and semantically coherent. As V-JEPA 2.1 uses 3D RoPE positional encoding, frequencies are interpolated according to the image resolution, enabling flexible processing across

resolutions. Figure 6 visualizes features from 64-frame video sequences, illustrating that dense video representations are also temporally consistent.

4.9 Familly of ViT-L and ViT-B distilled models. Table 8 shows that distilling ViT-G into ViT-L significantly improves performance over training ViT-L from scratch, nearly closing the gap with ViT-G. On SSv2 action recognition, accuracy increases from 74.2% to 76.5% (approaching 77.7% for ViT-G); ADE20k segmentation improves from 42.0 to 46.7 mIoU, and depth estimation improves from 2.490 to 2.461 RMSE. The distilled ViT-B also achieves performance comparable to ViT-L trained from scratch.

Table 8: Benefits of Distillation. Distilling the V-JEPA 2.1 ViT-G into smaller models (ViT-L, 300M; ViT-B, 80M) yields consistent improvements over training from scratch and approaches the performance of the ViT-G teacher.

Model	#Params	Distilled	SSv2	K400	IN1K	ADE20K	Citys.	VOC12	NYU (↓)	KITTI (↓)	DAVIS
ViT-G	2B		77.7	87.7	85.5	47.9	73.5	85.0	0.307	2.461	69.0
ViT-L	300M		74.2	84.3	81.8	42.0	66.7	79.6	0.381	2.914	64.5
ViT-L	300M	✓	76.5	86.9	84.9	46.7	71.0	83.9	0.336	2.490	68.7
ViT-B	80M	✓	74.5	85.1	84.0	42.0	64.5	78.0	0.408	2.850	67.0

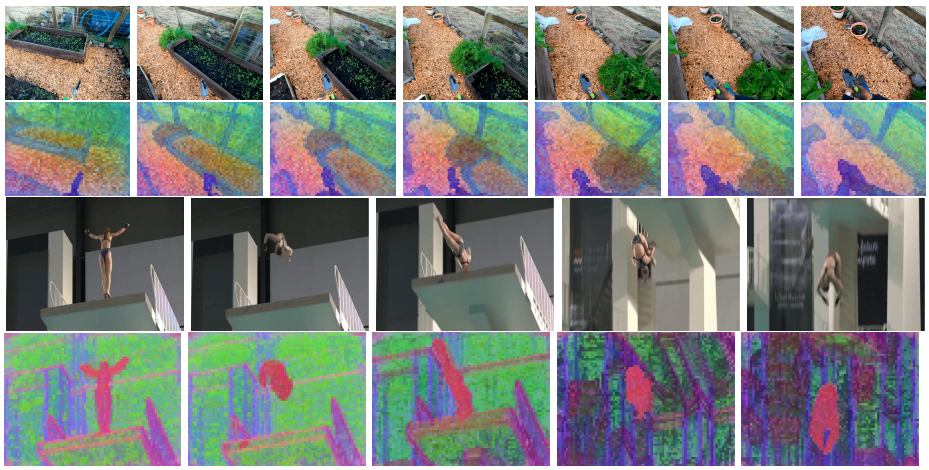


Fig. 6: V-JEPA 2.1 unlocks dense features from video. Qualitative results on Ego4D (1080px) and Diving48 (768px) illustrate the strong temporal consistency of the dense features produced by our ViT-G model, particularly on dynamic objects. We process entire video sequences of 64 frames using the 3D convolutional video tokenizer.

5 Conclusion

This work demonstrates that Joint-Embedding Predictive Architectures (JEPA) combined with spatio-temporal self-supervision can unlock dense local features with robust global understanding. Our core contribution—a weighted context loss paired with a multi-level predictor—enables supervision across intermediate layers, producing features that are spatially structured and temporally consistent. Together with modality-specific tokenizers, balanced image–video training, and scaling to ViT-G, this design also enables effective distillation to compact models. V-JEPA 2.1 achieves state-of-the-art results in action recognition and anticipation and sets new linear-probe benchmarks for depth estimation. These findings highlight the potential of predictive modeling to understand the physical world. Building on the world-modeling potential of V-JEPA, future work will investigate dense prediction capabilities within the world-model framework [12, 41], where world models with pixel-level precision will prove valuable for robotics and autonomous agents, particularly in tasks requiring fine-grained manipulation and navigation in complex environments.

Bibliography

[1] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: Vivit: A video vision transformer. In: ICCV (2021)

[2] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)

[3] Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)

[4] Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022)

[5] Balestriero, R., LeCun, Y.: Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544 (2025)

[6] Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)

[7] Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15791–15801 (June 2025)

[8] 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)

[9] Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)

[10] Bardes, A., Ponce, J., LeCun, Y.: Vicregl: Self-supervised learning of local visual features. NeurIPS (2022)

[11] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML (2021)

[12] Bojanowski, F.B.M.S.B.T.V.K.F.M.Y.L.P.L.M.S.P.: Back to the features: Dino as a foundation for video world models. arXiv preprint arXiv:2507.19468 (2025)

[13] Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)

[14] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)

[15] Caron, M., Touvron, H., Misra, I., Jegou, H., Joulin, J.M.P.B.A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
[16] Carreira, J., Gokay, D., King, M., Zhang, C., Rocco, I., Mahendran, A., Keck, T.A., Heyward, J., Koppula, S., Pot, E., Erdogan, G., Hasson, Y., Yang, Y., Greff, K., Moing, G.L., Steenkiste, S.v., Zoran, D., Hudson, D.A., Vélez, P., Polanía, L., Friedman, L., Duvarney, C., Goroshin, R., Allen, K., Walker, J., Kabra, R., Aboussouan, E., Sun, J., Kipf, T., Doersch, C., Pătrăucean, V., Damen, D., Luc, P., Sajjadi, M.S.M., Zisserman, A.: Scaling 4d representations. arXiv preprint arXiv:2412.15212 (2024)
[17] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
[18] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
[19] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
[20] Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision (IJCV) (2022)
[21] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
[22] Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context. In: ICCV (2015)
[23] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
[24] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
[25] Fan, D., Tong, S., Zhu, J., Sinha, K., Liu, Z., Chen, X., Rabbat, M., Ballas, N., LeCun, Y., Bar, A., Xie, S.: Scaling language-free visual representation learning. arXiv preprint arXiv:2504.01017 (2025)
[26] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J.: Multiscale vision transformers. In: ICCV (2021)
[27] Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: NeurIPS (2022)
[28] Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., LeCun, Y.: Intuitive physics understanding emerges from self-supervised pretraining on natural videos. arXiv preprint arXiv:2502.11831 (2025)
[29] Geiger, A., Lenz, P., Stiller, C., Urtasun., R.: Vision meets robotics: The kitti dataset. In: The International Journal of Robotics Research (2013)
[30] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)

[31] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)

[32] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

[33] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)

[34] Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3) (2018)

[35] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)

[36] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

[37] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[38] Hénaff, O.J., Koppula, S., Alayrac, J.B., van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. arXiv preprint arXiv:2103.10957 (2021)

[39] Jabri, A., Owens, A., Efros, A.: Space-time correspondence as a contrastive random walk. Advances in neural information processing systems **33**, 19545–19560 (2020)

[40] Jayaraman, D., Grauman, K.: Learning image representations tied to egomotion. In: ICCV (2015)

[41] Karypidis, E., Kakogeorgiou, I., Gidaris, S., , Komodakis., N.: Dino-foresight: Looking into the future with dino. arXiv preprint arXiv:2412.11673 (2024)

[42] Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., Fagan, P.D., Hejna, J., Itkina, M., Lepert, M., Ma, Y.J., Miller, P.T., Wu, J., Belkhale, S., Dass, S., Ha, H., Jain, A., Lee, A., Lee, Y., Memmel, M., Park, S., Radosavovic, I., Wang, K., Zhan, A., Black, K., Chi, C., Hatch, K.B., Lin, S., Lu, J., Mercat, J., Rehman, A., Sanketi, P.R., Sharma, A., Simpson, C., Vuong, Q., Walke, H.R., Wulfe, B., Xiao, T., Yang, J.H., Yavary, A., Zhao, T.Z., Agia, C., Baijal, R., Castro, M.G., Chen, D., Chen, Q., Chung, T., Drake, J., Foster, E.P., Gao, J., Guizilini, V., Herrera, D.A., Heo, M., Hsu, K., Hu, J., Irshad, M.Z., Jackson, D., Le, C., Li, Y., Lin, K., Lin, R., Ma, Z., Maddukuri, A., Mirchandani, S., Morton, D., Nguyen, T., O'Neill, A., Scalise, R., Seale, D., Son, V., Tian, S., Tran, E., Wang, A.E., Wu, Y., Xie, A., Yang, J., Yin, P., Zhang, Y., Bastani, O., Berseth, G., Bohg, J., Goldberg, K., Gupta, A., Gupta, A., Jayaraman, D., Lim, J.J., Malik, J.,

Martín-Martín, R., Ramamoorthy, S., Sadigh, D., Song, S., Wu, J., Yip, M.C., Zhu, Y., Kollar, T., Levine, S., Finn, C.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945 (2024)
[43] LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
[44] Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
[45] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
[46] Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: Unsupervised learning using temporal order verification. In: ECCV (2016)
[47] Mittal, H., Agarwal, N., Lo, S.Y., Lee, K.: Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18580–18590 (2024)
[48] Mo, S., Tong, S.: Connecting joint-embedding predictive architecture with contrastive self-supervised learning. Advances in neural information processing systems **37**, 2348–2377 (2024)
[49] Mur-Labadia, L., Martinez-Cantin, R., Guerrero, J.J., Farinella, G.M., Furnari, A.: Aff-ttention! affordances and attention models for short-term object interaction anticipation. In: European Conference on Computer Vision. pp. 167–184. Springer (2024)
[50] Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
[51] 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)
[52] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
[53] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
[54] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
[55] Ragusa, F., Farinella, G.M., Furnari, A.: Stillfast: An end-to-end approach for short-term object interaction anticipation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2023)
[56] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
[57] Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: CVPR (2024)

[58] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
[59] Roy, D., Rajendiran, R., Fernando, B.: Interaction region visual transformer for egocentric action anticipation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6740–6750 (2024)
[60] Silberman, N., Hoiem, D., Kohli, P., Fergus., R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
[61] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
[62] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
[63] Sutton, R.S.: An adaptive network that constructs and uses and internal model of its world. Cognition and Brain Theory **4**(3), 217–246 (1981)
[64] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
[65] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
[66] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: Look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
[67] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
[68] 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)
[69] Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
[70] Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European Conference on Computer Vision. pp. 396–416. Springer (2024)
[71] Xie, J., Zhan, X., Liu, Z., Ong, Y.S., Loy, C.C.: Unsupervised object-level representation learning from scene images. In: NeurIPS (2021)
[72] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. arXiv preprint arXiv:2111.09886 (2021)
[73] Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 585–601 (2018)

[74] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
[75] Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., Choi, Y.: Merlot reserve: Neural script knowledge through vision and language and sound. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16375–16387 (2022)
[76] Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
[77] Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024)
[78] Zhou, B., andXavier Puig, H.Z., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
[79] 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)