MV-JEPA: POST-TRAINING VIDEO ENCODERS FOR CROSS-VIEW UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

JEPA-based video encoders learn strong representations for video and image understanding through prediction in latent space. However, their standard pretraining predicts features within individual images or video clips, without explicitly exploiting paired camera views of the same object or moment. Consequently, the learned representations may not reliably support matching across different camera viewpoints. To solve this problem, we introduce multi-view predictive post-training to strengthen cross-view instance and temporal correspondence while retaining independent single-view inference. A shared encoder processes masked context views independently, and a training-time predictor jointly predicts latent tokens for masked context regions and a fully held-out view. We alternate cross-view prediction with single-view masked prediction on general image and video data to help retain the encoder's pretrained capabilities. After only 9,000 post-training updates on a single GPU, MV-JEPA improves cross-view instance retrieval and coarse cross-camera moment retrieval on various datasets, with further gains in zero-shot action-phase retrieval and frozen-feature classification on AE2. Frozen-probe evaluations further indicate that post-training largely preserves motion recognition, image recognition, and dense semantic segmentation performance on the evaluated benchmarks. Together, these results demonstrate that multi-view predictive post-training can strengthen instance and coarse temporal correspondence in JEPA video encoders without modifying the encoder architecture or single-view inference interface.

1 INTRODUCTION

Video encoders based on the Joint-Embedding Predictive Architecture (JEPA) learn visual representations through large-scale self-supervised learning (Bardes et al., 2024). V-JEPA 2, for example, achieves strong visual recognition performance using frozen encoder features (Assran et al., 2025). However, its pretraining objective predicts latent features within individual images or video clips, without explicitly exploiting paired camera views of the same object or moment. We therefore study short multi-view post-training to strengthen cross-view correspondence while preserving the general visual capabilities acquired during pretraining.

Cross-view correspondence requires representations to preserve identity despite changes in appearance. Different cameras can observe the same object from opposite sides or record the same activity at the same time, yet produce substantially different images. Useful features should match these observations while distinguishing other instances or moments. We therefore focus on viewpoint robustness for instance and temporal correspondence, with these distinctions retained in independently encoded features.

Prior work has explored cross-view representation learning through both contrastive and predictive objectives. Time-Contrastive Networks adapt an ImageNet-initialized visual model using synchronized multi-view contrastive learning (Sermanet et al., 2017), but their evaluation focuses on correspondence, task-specific attributes, and imitation without examining recognition retention after adaptation. BYOV combines self-view and cross-view masked prediction to learn temporal representations on top of a frozen pretrained image encoder (Park et al., 2025), leaving the pretrained visual backbone itself unchanged during cross-view training. Seq-JEPA learns invariant and equivariant representations through transformation-conditioned sequential prediction (Ghaemi et al., 2025). Its experiments study

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

self-supervised representation learning from scratch, without examining post-training adaptation of an already pretrained backbone. Its original formulation also uses explicit relative transformation parameters as conditioning inputs during training. Building on these advances, we study whether short post-training can strengthen cross-view correspondence in a pretrained JEPA video encoder while preserving its general recognition capabilities.

We introduce MV-JEPA, a multi-view predictive post-training recipe that directly updates the pretrained encoder. A shared encoder independently processes masked context views, and a joint predictor reconstructs masked context features and predicts the features of a fully held-out view. The held-out observation is used only to construct training targets, requiring its features to be inferred from other views without explicit inter-view transformation parameters. We further alternate this objective with single-view masked-prediction replay to help retain pretrained recognition capabilities.

Our contributions are:

1. **Short multi-view predictive post-training.** MV-JEPA strengthens pretrained encoders with only 9,000 updates on a single GPU, preserving their original single-view inference interface.
2. **Transferable cross-view understanding.** Improved instance and moment correspondence transfers across datasets and JEPA backbones, with additional benefits for ego–exo action understanding.
3. **Preserved general visual capabilities.** Cross-view gains coexist with largely retained motion recognition, image recognition, and dense semantic segmentation performance.

After 9,000 post-training updates, MV-JEPA (V-JEPA2-L) improves ABO instance retrieval by 9.4 percentage points and moment retrieval on unseen Ego-Exo4D captures by 14.6 points. The gains transfer to datasets not used for post-training and to V-JEPA2.1-g (Mur-Labadia et al., 2026). On AE2 (Xue & Grauman, 2023), zero-shot retrieval improves across all four actions, while classification and temporal ordering improve on three. Frozen probes show largely retained recognition performance on SSv2, Diving48, Jester, and ImageNet-1K, as well as comparable dense semantic segmentation performance on ADE20K.

2 Related Work

Cross-view correspondence learning. Prior work has studied correspondence through metric learning, temporal alignment, and predictive objectives. TCN learns embeddings from synchronized camera views and temporal negatives (Sermanet et al., 2017), while TCC learns temporal alignment through cycle consistency (Dwibedi et al., 2019). AE2 aligns unpaired ego–exo videos through object-centric encoding and a contrastive temporal-alignment objective (Xue & Grauman, 2023). Among predictive approaches, CroCo reconstructs masked pixels from visible image regions and an additional view of the same scene (Weinzaepfel et al., 2022). BYOV combines self-view and cross-view prediction of frame embeddings on top of a frozen image encoder (Park et al., 2025). Seq-JEPA studies transformation-conditioned sequential prediction for invariant and equivariant representations in a from-scratch training setting (Ghaemi et al., 2025). Our recipe directly updates a pretrained JEPA video backbone through joint prediction of masked context regions and a held-out view, while retaining independent single-view inference.

Post-training of visual representations. Targeted additional training can improve the capabilities of pretrained visual encoders. TimeTuning adapts image representations using temporal-alignment clustering on unlabeled videos (Salehi et al., 2023), and NeCo improves dense features through patch-neighbor ordering consistency (Pariza et al., 2025). MoSiC uses tracked motion trajectories to enforce temporal cluster consistency (Salehi et al., 2025), while 3DPoV uses tracked patches and ordering consistency to improve dense correspondence and geometric features (Simion et al., 2026). Adaptation also raises questions of retention: WiSE-FT interpolates pretrained and fine-tuned weights to balance adaptation with robustness under distribution shift (Wortsman et al., 2022). MoSiC additionally measures image classification performance alongside dense-task gains (Salehi et al., 2025). We study this adaptation problem in JEPA video encoders through masked-context and held-out-view latent prediction, interleaved with single-view replay, and evaluate instance and moment correspondence alongside recognition retention.

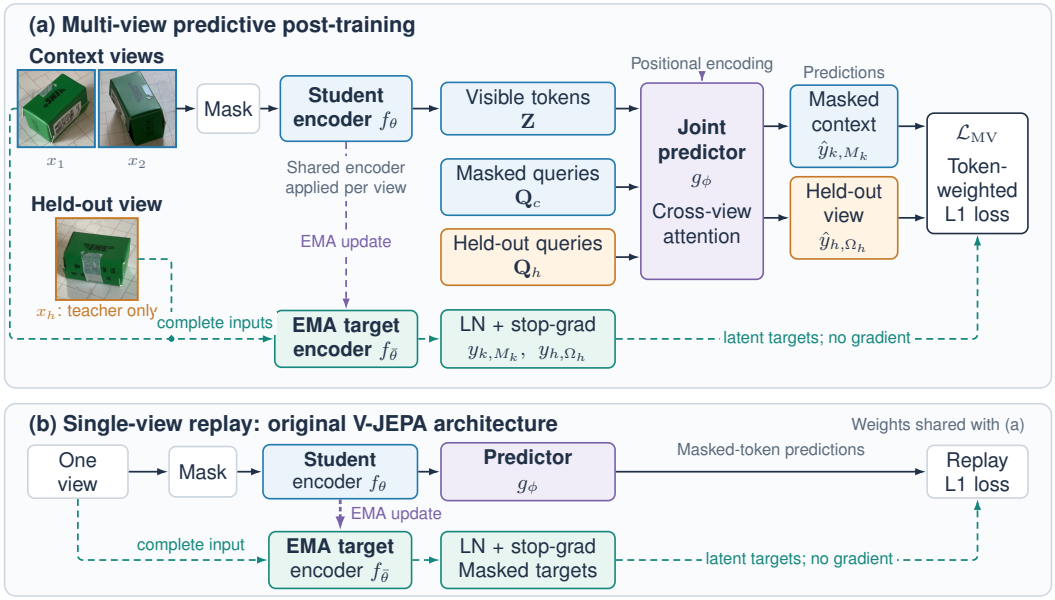

Figure 1: MV-JEPA overview. (a) A joint predictor reconstructs masked context features and predicts held-out-view features from independently encoded context views. (b) Single-view replay follows the original V-JEPA masked-prediction architecture and shares encoder and predictor weights with the multi-view branch. Each epoch contains 152 multi-view updates and 148 replay updates.

3 Multi-view Predictive Post-training

MV-JEPA adapts pretrained video encoders for cross-view understanding through short multi-view predictive post-training. It combines prediction of a held-out view from multiple context views with single-view replay of the original masked-prediction task. Figure 1 illustrates the framework.

Training data and preprocessing. We use two multi-view datasets for post-training. MVImgNet2.0 (Han et al., 2024) contains images of everyday objects captured from different viewpoints, extracted from short videos recorded around each object. Ego-Exo4D (Grauman et al., 2024) contains synchronized egocentric and exocentric recordings of skilled human activities, including cooking, sports, and music.

We sample from the two datasets with equal probability. Each training example consists of K context views, $\{x_k\}_{k=1}^{K}$, and one additional held-out view x_h. For MVImgNet2.0, the context comprises two to six images of the same object from different viewpoints. For Ego-Exo4D, it comprises up to three synchronized clips covering the same temporal interval from different viewpoints.

For MVImgNet2.0, each image is preprocessed using a square crop derived from its foreground mask. A target foreground-area ratio in $[0.35, 0.55]$ is sampled once per instance and shared across its views. Crop sizes adapt to each view while preserving the full foreground bounding box and retaining local background. The crops are resized to the input resolution (Figure 2), emphasizing the shared object in multi-view prediction. All matched training controls use the same preprocessed data.

Context encoding and latent targets. Our main implementation initializes the encoder and predictor from released V-JEPA2-L weights and retains its masked encoding and EMA target construction (Assran et al., 2025). As illustrated in Figure 1, f_θ denotes the student encoder and $f_{\bar{\theta}}$ its EMA target encoder.

For context view x_k, let V_k and M_k denote the visible and masked token positions, respectively. The student independently encodes each view's visible tokens to obtain context features z_k. The teacher

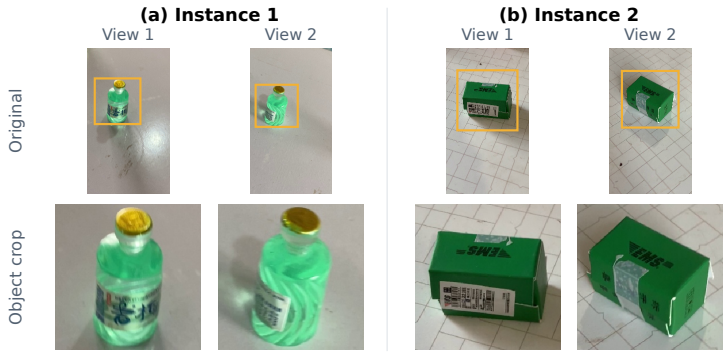

Figure 2: MVImgNet2.0 preprocessing. Orange boxes mark the square crops in the original views; the resulting 256×256 training images are shown below.

processes the complete context and held-out views to produce target features y_k and y_h:

$$\begin{aligned} z_k &= f_\theta(x_k; V_k), \\ y_k &= \mathrm{sg}[\mathrm{LN}(f_{\bar{\theta}}(x_k))], \\ y_h &= \mathrm{sg}[\mathrm{LN}(f_{\bar{\theta}}(x_h))]. \end{aligned} \tag{1}$$

Here, LN applies LayerNorm to each token, and sg stops gradients through the targets. We supervise masked context positions M_k and all held-out positions, denoted by Ω_h.

Context masking uses eight blocks per view at nominal spatial scale 0.15, following the V-JEPA2 block-mask sampler. The held-out observation enters only the teacher: neither the student nor the predictor receives its pixels or encoded features.

Joint multi-view prediction. The predictor combines the visible features from all context views to reconstruct masked context features and predict the held-out view.

For context view k, let $z_{k,p}$ denote the student feature at visible token position p. A shared projection P maps these features to the predictor width. We use two learned mask embeddings, m_c and m_h, to construct queries for masked-context and held-out prediction, respectively. Each query requests a feature at a position where no visual token is provided to the predictor.

We use learnable view embeddings to distinguish tokens from different views. For each example, we randomly assign distinct view IDs to the context views. Let s_k denote the embedding associated with the ID assigned to context view k, and let s_h denote a separate embedding reserved for the held-out view. Visible tokens and masked queries from the same context view share s_k. We also retain the original V-JEPA positional encoding to identify spatial and temporal locations within each view.

We denote the resulting visible-token inputs by $u_{k,p}$, masked-context queries by $q_{k,p}$, and held-out queries by $q_{h,p}$:

$$\begin{aligned} u_{k,p} &= P(z_{k,p}) + s_k, && p \in V_k, \\ q_{k,p} &= m_c + s_k, && p \in M_k, \\ q_{h,p} &= m_h + s_h, && p \in \Omega_h. \end{aligned} \tag{2}$$

Let $\mathbf{Z}$ collect all visible-token inputs $u_{k,p}$, $\mathbf{Q}_c$ all masked-context queries $q_{k,p}$, and $\mathbf{Q}_h$ all held-out queries $q_{h,p}$. The joint predictor g_ϕ, with parameters ϕ, processes their concatenation through self-attention. Let $\hat{y}_{k,p}$ and $\hat{y}_{h,p}$ denote its predicted features at context and held-out query positions:

$$\left(\{\hat{y}_{k,M_k}\}_{k=1}^{K}, \hat{y}_{h,\Omega_h}\right) = g_\phi\left(\mathrm{concat}[\mathbf{Z}, \mathbf{Q}_c, \mathbf{Q}_h]\right). \tag{3}$$

Here, $\hat{y}_{k,M_k}$ collects predictions at the masked positions of context view k, while $\hat{y}_{h,\Omega_h}$ collects predictions at every held-out position. Each query can attend to visible features from all context views, allowing both prediction tasks to use cross-view information.

Multi-view prediction objective. Using the V-JEPA2 token prediction error $e(\hat{y}, y) = D^{-1}\|\hat{y}-y\|_1$ for D-dimensional features, we supervise masked context positions and all held-out positions:

$$\mathcal{L}_{\mathrm{MV}} = \frac{\lambda_c \sum_{k=1}^{K} \sum_{p \in M_k} e(\hat{y}_{k,p}, y_{k,p}) + \lambda_h \sum_{p \in \Omega_h} e(\hat{y}_{h,p}, y_{h,p})}{\lambda_c \sum_{k=1}^{K} |M_k| + \lambda_h |\Omega_h|}. \tag{4}$$

We set $\lambda_c = \lambda_h = 1$, giving equal weight to each supervised token, and average the normalized per-example losses over the batch.

Masked-context prediction preserves the within-view reconstruction task while allowing additional views to provide context. Held-out prediction requires the context representations to support inference of another view's token-level features.

Replay and training configuration. To help retain general recognition capabilities, we interleave multi-view updates with single-view masked-prediction replay from ImageNet (Deng et al., 2009), Kinetics-700 (Carreira et al., 2019), and Something-Something-v2 (Goyal et al., 2017; Mahdisoltani et al., 2018). Replay shares the encoder and predictor with the multi-view branch and follows the original V-JEPA objective and two mask configurations. Each epoch contains a shuffled schedule of 152 multi-view and 148 replay updates. The EMA teacher is updated after each optimizer step with momentum 0.99925.

For V-JEPA2-L, video inputs consist of 16 sampled frames at 256×256 resolution. For image inputs, each image is resized to the same resolution and repeated to form a 16-frame clip. Multi-view and replay batch sizes are 48 and 24, respectively. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.04 and a post-training budget of 30 epochs, corresponding to 9,000 optimizer updates. Post-training V-JEPA2-L takes approximately 18 hours on a single NVIDIA RTX PRO 6000 Blackwell GPU.

Transfer to V-JEPA2.1-g. We apply the same multi-view objective to V-JEPA2.1-g, initializing the student and EMA teacher from its released EMA encoder weights. During post-training, both multi-view prediction and replay are supervised using 5,632-dimensional targets. These targets concatenate the EMA teacher's features from blocks 10, 20, 30, and 40, with each layer normalized separately. Student features from the same four layers form the predictor inputs. For evaluation, we adopt the final-layer feature readout used in V-JEPA2.1 (Mur-Labadia et al., 2026), extracting only the frozen EMA encoder's final-layer, 1,408-dimensional features.

Images use the native single-frame tokenizer, and image and video batches are processed in separate updates. Replay follows V-JEPA2.1 by combining masked-token prediction with a distance-weighted L1 loss on visible context tokens. Within the visible-token loss, tokens closer to masked positions receive greater weight. This loss term is scaled by 0.7 for images and 0.5 for videos.

4 Experiments

Models and shared setup. We primarily evaluate MV-JEPA using V-JEPA2-L. We additionally evaluate V-JEPA2.1-g to assess transfer of cross-view correspondence gains and retention of semantic segmentation performance. All post-trained models and training controls use their frozen EMA encoders from the final epoch-30 checkpoint, corresponding to 9,000 optimizer updates.

4.1 Cross-view instance retrieval

We evaluate cross-view instance recognition using frozen encoders and independently extracted features. For MVImgNet2.0, we use 5,700 query–gallery pairs across 285 categories, with 20 candidate instances per category; evaluated instances are excluded from post-training. For ABO (Collins et al., 2022), we use 7,460 spin-view pairs in 373 groups of 20 products, with product-type-matched candidates and white square padding. SOP (Song et al., 2016) uses its official test set of 60,502 images and 11,316 products, with each image querying the full gallery excluding itself. We report Recall@1 under one fixed protocol per dataset.

As shown in Table 1, MV-JEPA (V-JEPA2-L) improves Recall@1 over its released initialization by 15.81, 9.36, and 3.64 percentage points on MVImgNet2.0, ABO, and SOP, respectively. MV-JEPA (V-JEPA2.1-g) achieves corresponding gains of 11.58, 5.11, and 2.3 points.

Table 1: Cross-view instance retrieval with frozen encoders, Recall@1 (%).

Frozen encoder	MVImgNet2.0	ABO	SOP
V-JEPA2-L	50.58	52.82	43.97
MV-JEPA (V-JEPA2-L)	**66.39**	**62.18**	**47.61**
V-JEPA2.1-g	61.84	46.70	45.93
MV-JEPA (V-JEPA2.1-g)	**73.42**	**51.81**	**48.23**

Table 2: Cross-camera moment retrieval with frozen encoders, Top-1 accuracy (%) over six candidate windows.

Frozen encoder	Ego-Exo4D	Assembly101
V-JEPA2-L	30.11	31.37
MV-JEPA (V-JEPA2-L)	**44.75**	**39.70**
V-JEPA2.1-g	29.53	37.51
MV-JEPA (V-JEPA2.1-g)	**41.82**	**46.67**



Figure 3: Selected ABO retrieval examples with V-JEPA2-L and MV-JEPA (V-JEPA2-L). Each query uses its original gallery of 20 same-category products. Green and red borders indicate correct and incorrect top-1 product identities, respectively.

These results demonstrate improved cross-view instance discrimination, including among objects from the same category. Figure 3 illustrates this ability on ABO: MV-JEPA retrieves the same product from another viewpoint, while the released encoder selects a different product from the same category.

The gains on ABO and SOP show that this capability transfers to datasets not used for post-training. Consistent improvements for V-JEPA2-L and V-JEPA2.1-g further support the applicability of the post-training recipe across the two evaluated backbones.

4.2 Cross-camera moment retrieval

We evaluate cross-camera moment retrieval on two synchronized multi-view video datasets. Ego-Exo4D contains skilled human activities, such as sports and music, recorded from first- and third-person perspectives. We use four exocentric cameras for each of 81 takes from 71 captures excluded from post-training. Assembly101 (Sener et al., 2022) records people assembling and disassembling toy vehicles from multiple viewpoints. We use 360 recordings with eight static cameras per recording; this dataset is not used for post-training.

For each execution, we divide the valid time span into six equal intervals and sample a 16-frame clip at 4 fps around each interval's center, yielding six non-overlapping clips per camera. Each clip serves as a query to retrieve its synchronized counterpart among the six clips from another camera, using cosine similarity between independently extracted features. We evaluate all ordered camera pairs and compute the percentage of correct retrievals within each take or recording. The final Top-1 score averages these accuracies, giving equal weight to each take or recording.

As shown in Table 2, MV-JEPA (V-JEPA2-L) improves Top-1 accuracy over its released initialization by 14.64 percentage points on Ego-Exo4D and 8.33 points on Assembly101. MV-JEPA (V-JEPA2.1-g) achieves corresponding gains of 12.29 and 9.16 points.

MV-JEPA more accurately matches the same moment of an activity across viewpoints, while distinguishing it from other moments within the same execution. The gains on Assembly101 show transfer to a dataset not used for post-training, while consistent improvements on both JEPA backbones support the transferability of our post-training recipe. Together, these results demonstrate improved temporal discrimination across viewpoints, supporting stronger cross-view understanding of dynamic scenes.

Table 3: AE2 action understanding with frozen encoders. Classification and retrieval report weighted F1 (%) and mAP@10 (%), respectively.

		Classification (F1)		Retrieval (mAP@10)		Kendall's
Task	Frozen encoder	Regular	Exo→Ego	Regular	Exo→Ego	$\tau \uparrow$
Break Eggs	V-JEPA2-L	63.80	20.17	59.48	40.93	0.1123
	MV-JEPA (V-JEPA2-L)	62.70	19.27	60.07	43.72	0.1024
Pour Milk	V-JEPA2-L	68.06	40.09	72.68	53.95	0.2598
	MV-JEPA (V-JEPA2-L)	71.19	54.14	75.24	60.05	0.3154
Pour Liquid	V-JEPA2-L	71.43	38.71	70.41	53.06	0.1051
	MV-JEPA (V-JEPA2-L)	75.21	57.62	71.58	58.19	0.1093
Tennis Forehand	V-JEPA2-L	92.16	63.79	90.15	63.53	0.4833
	MV-JEPA (V-JEPA2-L)	93.68	66.30	91.74	63.56	0.5391

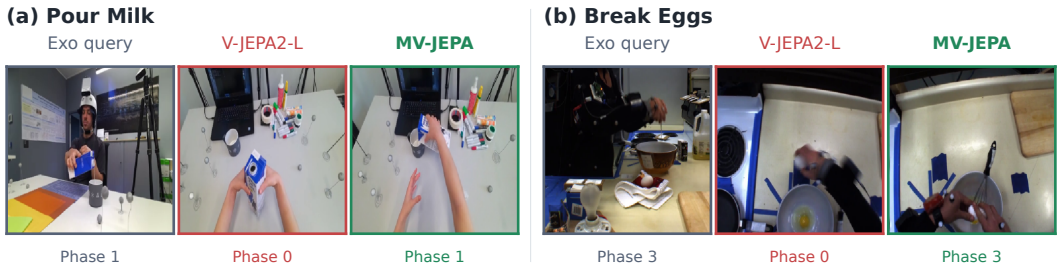

Figure 4: Selected Exo→Ego retrieval examples on AE2. Each panel shows a query frame and the top-1 retrieved frames from V-JEPA2-L and MV-JEPA. Green and red borders indicate matching and mismatched action phases, respectively.

4.3 Ego–Exo action understanding on AE2

AE2 contains egocentric and exocentric videos of Break Eggs, Pour Milk, Pour Liquid, and Tennis Forehand, with frame-level action-phase annotations. We compare frozen V-JEPA2-L and MV-JEPA (V-JEPA2-L) using its official training and test splits. AE2 is not used in MV-post-training.

Both encoders process 16-frame clips at 256×256 resolution under an offline protocol. Frame features are extracted from the corresponding two-frame tubelet positions, retaining the original phase labels. Positions whose constituent frames have different labels are excluded identically for both models during probe training and evaluation.

Classification trains an RBF SVM with $C = 1$ and `gamma=scale`, reporting weighted F1. Zero-shot retrieval reports mAP@10 using the AE2 implementation, with matching phase labels defining relevance and the query video excluded from the gallery. Both tasks use Regular and Exo→Ego settings. Regular uses mixed ego/exo views. Exo→Ego classification trains on exocentric videos and tests on egocentric videos; retrieval uses exocentric queries and an egocentric gallery. Kendall's τ evaluates temporal-order preservation of nearest-neighbor frame matches across distinct videos.

Table 3 shows retrieval gains in both settings across all four actions. For Pour Milk and Pour Liquid, respectively, Exo→Ego mAP@10 increases from 53.95% to 60.05% and from 53.06% to 58.19%, while classification F1 increases from 40.09% to 54.14% and from 38.71% to 57.62%. Classification and Kendall's τ improve on three of the four actions.

Figure 4 shows successful Exo→Ego matches on Pour Milk and Break Eggs: MV-JEPA retrieves the query's action phase, while the released encoder selects a different phase. Together, these results support improved action-phase discrimination and matching across viewpoints and executions.

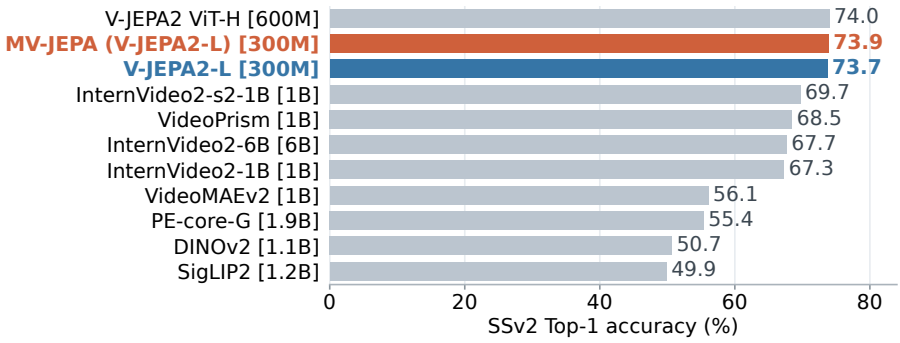

Figure 5: Frozen-probe recognition on SSv2, Top-1 (%). Selected references include VideoMAEv2 (Wang et al., 2023), InternVideo2 (Wang et al., 2024), VideoPrism (Zhao et al., 2024), DINOv2 (Oquab et al., 2023), PE (Bolya et al., 2025), SigLIP2 (Tschannen et al., 2025), and V-JEPA2 (Assran et al., 2025); scores follow Assran et al. (2025, Table 4). MV-JEPA uses a reduced 6-configuration search compared with the published V-JEPA2-L baseline.

4.4 General visual capabilities with frozen probes

We assess whether MV-post-training preserves general visual capabilities while improving cross-view understanding. Recognition evaluations use frozen V-JEPA2-L encoders with attentive probes, while semantic segmentation uses frozen V-JEPA2.1-g encoders with linear probes.

Motion understanding. We evaluate SSv2, Diving48 (Li et al., 2018), and Jester (Materzynska et al., 2019). For SSv2, we compare MV-JEPA with published reference results reproduced in Assran et al. (2025, Table 4). For Diving48 and Jester, both encoders use identical lightweight probe protocols. We use 4,077 Diving48 training videos and 29,641 Jester training videos, corresponding to 27.1% and 25.0% of their full training sets, respectively. Evaluation retains the complete 1,970-video Diving48 test set and 14,787-video Jester validation set.

As shown in Figure 5, MV-JEPA achieves 73.9% Top-1 accuracy on SSv2, compared with the published V-JEPA2-L result of 73.7%. We follow the V-JEPA2 frozen-probe framework with a reduced 6-configuration search, using a subset of the 20-configuration grid used for the published baseline.

Under the matched lightweight protocols (Table 4), Diving48 accuracy changes from 83.30% to 82.79%, while Jester accuracy increases from 96.84% to 96.98%. These modest changes indicate that motion recognition is largely retained alongside the gains in cross-view correspondence.

Appearance understanding. We evaluate ImageNet-1K using a class-balanced subset of 200,000 training images and the complete 50,000-image validation set. Inputs have resolution 256×256, with each image repeated to 16 frames. Both encoders use identical probe-training settings: 15 epochs and four hyperparameter configurations. We report the best validation accuracy among these configurations at the final probe epoch.

Table 4 shows that MV-JEPA achieves 76.57% Top-1 accuracy, compared with 76.74% for the released encoder. This small difference indicates that appearance recognition is also largely retained under the matched protocol.

Dense prediction. We evaluate semantic segmentation on ADE20K (Zhou et al., 2017) using its complete sets of 20,210 training and 2,000 validation images. For both encoders, we train a linear probe on frozen final-layer patch features for 40,000 updates, using 512×512 training crops and identical settings with a single fixed hyperparameter configuration. As shown in Figure 6, both encoders achieve closely matched performance throughout probe training. At the final update, MV-JEPA achieves 47.11 mIoU, compared with 47.23 for the released encoder. These results indicate that dense semantic segmentation performance is largely preserved, extending the retention evidence to dense prediction with V-JEPA2.1-g.

Table 4: Frozen-probe recognition, Top-1 (%). Both encoders use V-JEPA2-L (Assran et al., 2025) (300M) with matched lightweight protocols.

	Motion Understanding		Appearance Understanding
Frozen encoder	Diving48	Jester	ImageNet-1K
V-JEPA2-L	83.30	96.84	76.74
MV-JEPA (V-JEPA2-L)	82.79	96.98	76.57

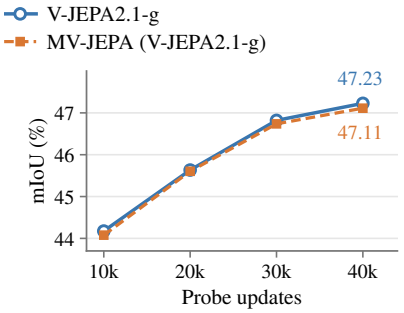

Figure 6: ADE20K linear-probe curves.

Table 5: Ablation of multi-view post-training on V-JEPA2-L. Scores are R@1 (%) for instance retrieval and Top-1 (%) for cross-camera moment retrieval, following the protocols in Tables 1 and 2.

Frozen encoder / variant	MVImgNet2.0	ABO	SOP	Ego-Exo4D
V-JEPA2-L	50.58	52.82	43.97	30.11
V-JEPA2-L + single-view post-training	50.19	50.63	43.77	29.97
MV-JEPA w/o held-out prediction	53.33	55.79	44.83	30.98
MV-JEPA (V-JEPA2-L)	**66.39**	**62.18**	**47.61**	**44.75**

4.5 Ablations and attribution

We compare MV-JEPA (V-JEPA2-L) with two controls to examine the contributions of multi-view context and held-out prediction. All three post-trained variants include the same replay branch, with identical replay data, masks, batch sizes, and update counts. V-JEPA2-L + single-view post-training uses one context view without held-out prediction. MV-JEPA w/o held-out prediction retains multi-view context but removes held-out queries and their supervision, including the associated prediction computation. All variants share the post-training data sources and 9,000-update budget.

Table 5 shows that V-JEPA2-L + single-view post-training does not improve over the released encoder on any of the four tasks. MV-JEPA w/o held-out prediction provides modest gains, reaching 53.33% on MVImgNet2.0, 55.79% on ABO, 44.83% on SOP, and 30.98% on Ego-Exo4D. The complete MV-JEPA model further improves these scores by 13.06, 6.39, 2.78, and 13.77 percentage points, respectively.

These comparisons show that additional training on the same data sources and access to multiple context views alone do not account for the full gains. The held-out prediction branch substantially improves both recognition of the same object across views and matching of synchronized activity moments across cameras. This supports held-out cross-view prediction as a key component of MV-JEPA's improved cross-view understanding.

5 Conclusion

We introduced MV-JEPA, a post-training recipe combining multi-view latent prediction with single-view replay. After 9,000 updates, it improves cross-view instance and coarse temporal correspondence across datasets and two JEPA backbones, with further benefits for ego–exo action understanding. Frozen-probe evaluations indicate largely preserved recognition and semantic segmentation performance on the evaluated benchmarks, while matched controls support the contribution of held-out-view prediction. These results demonstrate that targeted predictive post-training can strengthen cross-view understanding while retaining the original encoder architecture and independent single-view inference interface.

AI USE STATEMENT

Generative AI tools assisted with literature search, software implementation and debugging, and manuscript polishing. The authors are responsible for the final content of this work, including AI-assisted text, claims, code, and other artifacts.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception Encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. URL https://arxiv.org/abs/2504.13181.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. URL https://arxiv.org/abs/1907.06987.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Collins_ABO_Dataset_and_Benchmarks_for_Real-World_3D_Object_Understanding_CVPR_2022_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1801–1810, 2019. URL https://sites.google.com/view/temporal-cycle-consistency.

Hafez Ghaemi, Eilif Benjamin Muller, and Shahab Bakhtiari. seq-JEPA: Autoregressive predictive learning of invariant-equivariant world models. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=GKt3VRaCU1.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5842–5850, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Goyal_The_Something_Something_ICCV_2017_paper.html.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Grauman_Ego-Exo4D_Understanding_Skilled_Human_Activity_from_First-_and_Third-Person_Perspectives_CVPR_2024_paper.html.

Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. MVImgNet2.0: A larger-scale dataset of multi-view images. *ACM Transactions on Graphics*, 43(6), 2024. doi: 10.1145/3687973. URL https://doi.org/10.1145/3687973.

Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yingwei_Li_RESOUND_Towards_Action_ECCV_2018_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint arXiv:1804.09235*, 2018. doi: 10.48550/arXiv.1804.09235. URL https://arxiv.org/abs/1804.09235.

Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The Jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. URL https://openaccess.thecvf.com/content_ICCVW_2019/html/HANDS/Materzynska_The_Jester_Dataset_A_Large-Scale_Video_Dataset_of_Human_Gestures_ICCVW_2019_paper.html.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Valentinos Pariza, Mohammadreza Salehi, Gertjan J. Burghouts, Francesco Locatello, and Yuki M. Asano. Near, far: Patch-ordering enhances vision foundation models' scene understanding. In *International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fbc9981dd6316378aee7fd5975250f21-Abstract-Conference.html.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bootstrap your own views: Masked ego-exo modeling for fine-grained view-invariant video representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.19706.

Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, and Yuki M. Asano. Time does tell: Self-supervised time-tuning of dense image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Salehi_Time_Does_Tell_Self-Supervised_Time-Tuning_of_Dense_Image_Representations_ICCV_2023_paper.html.

Mohammadreza Salehi, Shashanka Venkataramanan, Ioana Simion, Efstratios Gavves, Cees G. M. Snoek, and Yuki M. Asano. MoSiC: Optimal-transport motion trajectory for dense self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6541–6551, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Salehi_MoSiC_Optimal-Transport_Motion_Trajectory_for_Dense_Self-Supervised_Learning_ICCV_2025_paper.html.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Sener_Assembly101_A_Large-Scale_Multi-View_Video_Dataset_for_Understanding_Procedural_Activities_CVPR_2022_paper.html.

Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. URL https://sermanet.github.io/tcn/.

Ioana Simion, Mohammadreza Salehi, Shashanka Venkataramanan, Cees G. M. Snoek, and Yuki M. Asano. 3DPoV: Improving 3d understanding via patch ordering on videos. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=CxcHPeXYWB.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://github.com/rksltnl/Deep-Metric-Learning-CVPR16.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_VideoMAE_V2_Scaling_Video_Masked_Autoencoders_With_Dual_Masking_CVPR_2023_paper.html.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. URL https://arxiv.org/abs/2403.15377.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2210.10716.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2109.01903.

Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/a845fdc3f87751710218718adb634fe7-Abstract-Conference.html.

Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. VideoPrism: A foundational visual encoder for video understanding. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.13217.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html.