AHJEPA: A SINGLE BRANCH FRAMEWORK THAT CAPTURES GRANULAR DETAILS FOR JEPA METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Regularization-based self-supervised learning methods enable stable training without additional heuristic[1] designs. However, such methods do not encourage learning of fine-grained relationships, which limits their applicability on downstream tasks. We mathematically characterize contextual prediction from incomplete observations and derive its implementation through a shared backbone, without a separately parameterized teacher or EMA updates. Our analysis interprets stop-gradient (SG) as a restriction on target-side adaptation, and an exploratory complementary-subset objective illustrates a possible implementation without SG. Inspired by previous masked-image-modeling (MIM) methods, we propose **AhJEPA**, a simple yet effective architecture that is broadly **A**pplicable to **h**euristic-free **JEPA** losses to encourage learning of fine-grained relationships of the input data while maintaining the training simplicity and stability. Similar to previous works, we add a reconstruction loss to force the model to learn the fine-grained details. Different from them, AhJEPA maintains a one-branch training paradigm and applies the reconstruction loss to CLS tokens rather than patch tokens. We retain SG in the final method for its stronger empirical performance. We find that this contextual learning approach is a more general way to improve JEPA methods with different special designs (*e.g.*, sparsity creation). This simple method consistently improves the performance of heuristic-free JEPA regularization methods across in-domain and out-of-domain classification, segmentation, depth estimation, and latent world model planning. Specifically, it helps the VISReg method achieve the best OOD accuracy and reduce the gap between heuristic-free methods and iBOT on fine-grained tasks. Also, it boosts the LeWorldModel performance to SOTA on OGBench-Cube and Two-Room datasets.

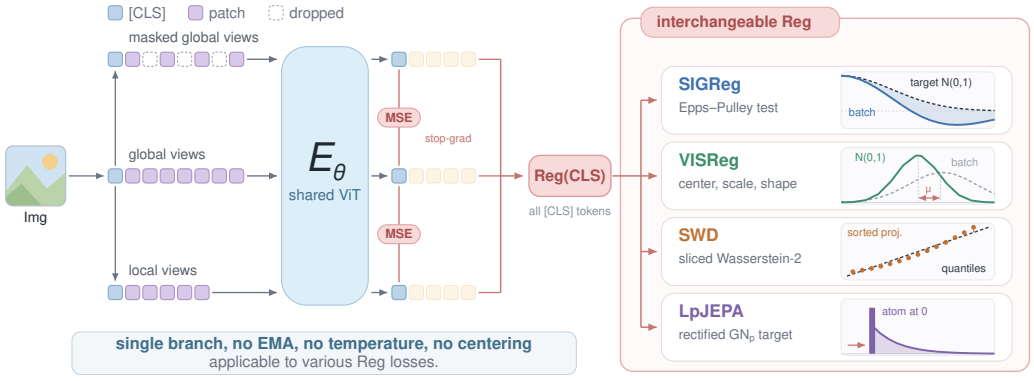

Figure 1: The architecture of AhJEPA is single-branch and applicable to existing JEPA regularization losses (Balestriero & LeCun, 2025; Wu et al., 2026; Bonneel et al., 2015; Kuang et al., 2026). With this simple architecture, AhJEPA improves the performance of the model trained with each on various tasks.

[1]Definition of heuristics in this paper: empirical techniques used for training stability.

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.

1 INTRODUCTION

Regularization-based JEPA methods (Bardes et al., 2022; Balestriero & LeCun, 2025; Kuang et al., 2026; Wu et al., 2026) alleviate the pain of tuning heuristics for training stability while achieving a strong downstream performance. This handy feature allows them to be applied to various domains.

However, these methods miss the contextual learning component, which has been shown to be an effective approach to capture granular details in SSL methods (Zhou et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). To fill this gap, we first look at the existing methods. After two major evolutions on DINO (Caron et al., 2021), DINOv3 (Siméoni et al., 2025) has provided a solid solution for learning granular details. It relies on the teacher-student architecture and a bag of heuristics (*e.g.*, stop gradient, temperature, and EMA), which makes it hard to apply in other domains. Hence, we seek a method that is broadly applicable to both current and future JEPA approaches without heavily relying on heuristics.

We first ask what contextual prediction requires a representation to encode, and whether these requirements depend on a teacher–student implementation. We formulate iBOT's masked prediction objective as contextual representation learning and show that a separately parameterized teacher or EMA updates is not mandatory. By decomposing the shared-backbone gradient, we identify that stop-gradient provides a mechanism to prevent the global feature from degrading into a masked global feature. This suggests a possibility to remove SG by adding explicit regularization. Our empirical results also support this possibility, motivating alternative ways to regulate joint adaptation. Nevertheless, we retain SG in our final method because it achieves the best performance.

In light of the above, we propose AhJEPA, a single-branch contextual learning method that improves the performance of JEPA methods that have different special designs (*e.g.*, sparsity creation). Similar to iBOT, it creates incomplete views for an image and aligns them to the complete view. Different from iBOT, AhJEPA directly drops the patch tokens and aligns views with CLS tokens. We choose this design for two reasons: 1) Patch token alignment hurts the embedding quality when the loss, *i.e.*, LpJEPA (Kuang et al., 2026), is designed to create sparse representations; 2) Adding learned masking tokens does not bring more benefits but a heavier compute burden for the CLS token reconstruction approach. We also control the number of augmentations to indicate that the improvements are from the AhJEPA design rather than additional views.

To comprehensively evaluate the efficacy of AhJEPA, we conducted experiments on 22 datasets, covering linear probe results on in-domain and OOD classification, segmentation, depth estimation, and the latent world model planning. AhJEPA increases the average accuracy of VISReg (Wu et al., 2026) by 1% and 0.29% on in-domain and OOD classification, respectively. This helps VISReg achieve the higher SOTA OOD accuracy and reduce the accuracy gap to iBOT on in-domain datasets. Regarding segmentation and depth estimation, AhJEPA increases the mIoU score of VISReg by at least 0.89% and improves NYU Depth v2 RMSE from 0.688 to 0.663. Lastly, we directly apply AhJEPA to the LeWorldModel (Maes et al., 2026) training pipeline. With a small tuning on learning rate and λ, it increases the success rate on OGBench-Cube and Two-Room by 4% and 11%, achieving the SOTA performance. All these results show that AhJEPA is helpful across the evaluated task families, which is more useful than the vanilla JEPA methods. Our contributions are:

- We mathematically characterize contextual representation learning through a reconstruction decomposition and local predictability bounds.
- We interpret stop-gradient as a form of gradient regularization that restricts direct target-side adaptation, and characterize the adaptation pathway it removes in shared-encoder contextual learning. This analysis motivates alternative ways to regulate joint adaptation.
- We propose AhJEPA, a single-branch training method that can be broadly applied to JEPA methods to improve the overall downstream performance.

2 RELATED WORK

Heuristic-free JEPA losses. Different from contrastive learning methods relying on negative samples to regulate the latent space, heuristic-free JEPA losses directly regulate the latent space with a regularization function. This simple and mathematically grounded approach removes the need to

add heuristics to ensure training stability. VICReg (Bardes et al., 2022) uses variance and covariance to maximize information in each dimension, but covariance only covers second-order statistics, failing to regulate the shape of the full distribution needed for stable training. SIGReg (Balestriero & LeCun, 2025) combines Epps-pulley (Epps & Pulley, 1983) and characteristics function of the Gaussian distribution to force the latent to an isotropic Gaussian. VISReg (Wu et al., 2026) fixes the gradient vanishing issue of SIGReg when embedding collapses by decomposing regularization into center, scale, and shape — maintaining the advantages of SIGReg while providing flexibility of latent space control. LpJEPA (Kuang et al., 2026) uses SWD (Bonneel et al., 2015) and a rectified generalized Gaussian distribution to achieve sparse embeddings. This paper uses SIGReg, VISReg, SWD, and LpJEPA to verify the applicability of AhJEPA.

Masked Image Modeling (MIM). Inspired by BERT (Devlin et al., 2019), MIM has become an important approach to visual representation learning, with prediction objectives defined over pixels (He et al., 2022; Xie et al., 2022) or patch-level representations (Zhou et al., 2021; Bao et al., 2022; Oquab et al., 2024; Siméoni et al., 2025). Closely related to our motivation, iBOT extends DINO by combining image-level self-distillation with masked patch prediction, using an EMA teacher as an online tokenizer to provide patch-level targets. This combination promotes local semantic learning and improves performance on dense prediction tasks. We study contextual prediction through a mathematical analysis of CLS reconstruction, characterizing the prediction requirements of different views and interpreting stop-gradient as gradient regularization that restricts target-side adaptation. This analysis motivates AhJEPA, a single-branch framework that enhances JEPA methods through CLS reconstruction, and an exploratory alternative without stop-gradient.

3 Contextual Learning with a Shared Encoder

We study contextual reconstruction as a means of extending distributionally regularized JEPA methods to fine-grained tasks. Starting from a formulation that includes iBOT, we characterize the prediction requirements of AhJEPA's CLS objective and derive how they are implemented through a shared backbone. We then examine how stop-gradient (SG) regulates target adaptation. Proofs and supplementary results are provided in Appendix A.1.

3.1 What contextual prediction requires

Masked image modeling (MIM) learns to predict information about a complete image from an incomplete observation. For an image X and a masking operator M, write $V = M(X)$. The contextual prediction objective is

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{X,M}\, \ell\big(q_\theta(V), T_\phi(X, M)\big), \tag{1}$$

where q_θ is the prediction and T_ϕ is the reconstruction target. iBOT (Zhou et al., 2021) is a patch-level instance. AhJEPA instead reconstructs the clean CLS embedding from dropped views and aligns local embeddings with their parent global embedding. These formulations differ in their targets and implementations, but all learn predictions from restricted visual evidence.

Let x_i, $i = 1, \ldots, G$, be the global views of an image. Each has K patch-token-dropped views $\widetilde{x}_{ik}$ and a local views x_{ij}^{loc}, sampled within its parent crop before view-specific augmentation. The embeddings produced by the shared encoder $F_\theta(\cdot)$ are written as

$$z_i = F_\theta(x_i), \qquad u_{ik} = F_\theta(\widetilde{x}_{ik}), \qquad v_{ij} = F_\theta(x_{ij}^{\text{loc}}) \in \mathbb{R}^d. \tag{2}$$

We analyze our final CLS objective with $\mathsf{D}(z, q) = \|z - q\|_2^2/d$ and $q = F_\theta(V)$ for either partial observation type. Expectations are over images and views at fixed parameters. We assume a continuous encoder with finite parameters on a compact image domain, ensuring finite second moments, and hold batch-normalization statistics fixed. For gradient identities, we additionally assume differentiability and interchangeability of differentiation and expectation.

Lemma 1 (Contextual reconstruction decomposition)**.** *Let z be a parent global embedding and $q = F_\theta(V)$ the embedding of a dropped or local observation V. Then*

$$\begin{aligned}\mathbb{E}\,\mathsf{D}(z, q) = \frac{1}{d}\mathbb{E}\|z - \mathbb{E}[z \mid V]\|_2^2 \\ + \frac{1}{d}\mathbb{E}\|q - \mathbb{E}[z \mid V]\|_2^2.\end{aligned} \tag{3}$$

The decomposition separates target variation unpredictable from the observation from prediction error relative to the conditional mean. For a fixed target, squared-error minimization favors this mean without requiring an explicit conditional-mean module. The identity also applies to the Patch candidate when V includes the masked input and queried position, averaging over masked positions used by the loss.

Dropped views require prediction from patches retained across the global field of view, whereas local views require prediction of global information from a restricted region. We next characterize local predictability within the CLS objective.

Theorem 1 (Local predictability of a nonconstant global representation)**.** *Let z be a parent global embedding, V_L an associated local observation, and $v = F_\theta(V_L)$. Suppose the total variance $\sigma_z^2 = \operatorname{tr}\operatorname{Cov}(z) > 0$. Define the fraction of global embedding variance predictable from local evidence as*

$$\eta_L(z) = \frac{\operatorname{tr}\operatorname{Cov}\big(\mathbb{E}[z \mid V_L]\big)}{\sigma_z^2}. \tag{4}$$

Then $0 \le \eta_L(z) \le 1$, and

$$\boxed{\mathcal{L}_L := \mathbb{E}\,\mathsf{D}(z, v) \ge \frac{\sigma_z^2}{d}(1 - \eta_L(z)).} \tag{5}$$

Under $\operatorname{Cov}(z) = I_d$, $\eta_L(z) \ge \max\{0, 1 - \mathcal{L}_L\}$.

At fixed nonzero global variance, low local alignment error requires a large fraction of global representation variance to be predictable from local evidence. Corollary 1 in Appendix A.1 extends this requirement to the combined dropped- and local-view objective.

3.2 Realizing contextual learning with one-branch training

The preceding prediction requirements do not require separate teacher and student networks. We now examine updates through a shared encoder F_θ. Define the per-parent squared alignment loss, differentiated through both predictions and targets, as

$$\ell_i^{\text{full}} = \frac{\sum_{k=1}^{K} \mathsf{D}(z_i, u_{ik}) + \sum_{j=1}^{a} \mathsf{D}(z_i, v_{ij})}{a + K}, \qquad \widehat{\mathcal{L}}_{\text{ctx}}^{\text{full}} = \frac{1}{G}\sum_i \ell_i^{\text{full}}. \tag{6}$$

Proposition 1 (Shared-backbone contextual gradients)**.** *Let $J_w = \partial w / \partial \theta$. Full backpropagation gives*

$$\begin{aligned} \nabla_\theta \ell_i^{\text{full}} = \frac{2}{d(a+K)} \Bigg[& \sum_k J_{u_{ik}}^\top (u_{ik} - z_i) \\ & + \sum_j J_{v_{ij}}^\top (v_{ij} - z_i) \\ & + J_{z_i}^\top \left(\sum_k (z_i - u_{ik}) + \sum_j (z_i - v_{ij}) \right) \Bigg]. \end{aligned} \tag{7}$$

The first two terms differentiate through dropped and local predictions; the last differentiates through their common target.

These coupled gradients train prediction from restricted evidence and adapt the common target through shared parameters. Neither a separately parameterized teacher, EMA, nor SG is needed to define this objective or compute its gradients.

3.3 Stop-gradient and target adaptation

The base regularizer constrains output statistics but allows the represented content to change. We next identify the part of target adaptation that can reduce the information to be inferred from an incomplete observation.

Proposition 2 (Stop-gradient restricts target adaptation). *For a dropped-view pair (z, u) with observation V, write $m_\theta(V) = \mathbb{E}[z_\theta \mid V]$ and*

$$\varepsilon_V(\theta) = \mathbb{E}\|z_\theta - m_\theta(V)\|_2^2. \tag{8}$$

Its expected target-side gradient decomposes as

$$\begin{aligned} g_{\text{target}}(V) &:= \frac{2}{d}\mathbb{E}[J_z^\top (z-u)] \\ &= \frac{1}{d}\nabla_\theta \varepsilon_V + \frac{2}{d}\mathbb{E}[J_z^\top (m_\theta(V) - u)]. \end{aligned} \tag{9}$$

Applying SG to the clean targets of dropped-view reconstruction in Eq. 13 removes precisely

$$g_{D,\text{target}} = \frac{2}{dG(a+K)} \sum_{i,k} J_{z_i}^\top (z_i - u_{ik}) \tag{10}$$

from $\nabla_\theta \widehat{\mathcal{L}}_{\text{ctx}}^{\text{full}}$ and leaves the local-target, global, and regularization contributions unchanged.

SG thus acts as gradient regularization on target adaptation. In isolation, descent using the removed conditional-variance component locally reduces variation unpredictable from dropped views, potentially weakening supervision if that variation encodes useful fine-grained relationships. SG blocks this route without freezing the shared encoder. The net effect depends on all gradient contributions, so target adaptation is not necessarily harmful.

Implications for fine-grained understanding. AhJEPA combines contextual prediction from dropped views with local predictability, encouraging relationships between local evidence and broader image content. SG restricts a direct route by which the clean target could shed information difficult to predict under dropping. These results explain the prediction requirements and optimization choices — improvements in segmentation and depth estimation are downstream benefits of the learned relationships. We also explore the possibility of removing SG by adding additional regularization terms, and details are in Appendix A.1.3.

4 AhJEPA: Applicable to heuristic-free JEPA methods

Building on the shared-encoder analysis, we now develop a contextual learning method that broadly applies to heuristic-free JEPA losses.

4.1 A dedicated MIM design for one-branch training

Loss candidates. SIGReg (Balestriero & LeCun, 2025) uses characteristic functions to match the latent distribution to an isotropic Gaussian; VISReg (Wu et al., 2026) separates center, scale, and shape regularization; SWD (Bonneel et al., 2015) performs distribution matching through one-dimensional optimal transport projections; and LpJEPA (Kuang et al., 2026) promotes sparse embeddings through a rectified reference distribution. We use these four regularizers to evaluate the applicability of AhJEPA across different latent-space constraints. To allocate our computational budget, we first compare their costs at different scales (Appendix Figure 6). Since SWD is the most efficient in this comparison, we use it for the initial design study. Following the invariance–regularization formulation of SIGReg and VISReg, the base objective is

$$\mathcal{L}_{\text{base}} = (1-\lambda)\mathcal{L}_{\text{inv}} + \lambda \mathcal{R}, \tag{11}$$

where $\mathcal{R}$ denotes the selected regularizer.

MIM approaches. iBOT (Zhou et al., 2021) performs masked prediction with a teacher–student architecture. To adapt this principle to a shared backbone, we compare three patch-prediction head designs: a shared MLP, a separate MLP with EMA updates, and a separate MLP with weight decay disabled. We additionally evaluate two reconstruction targets: 1) **CLS**, which drops input patch tokens and predicts the clean CLS projection through the shared MLP, and 2) **Patch**, which uses a JEPA-type (LeCun et al., 2022) asymmetric MLP to predict clean patch features at masked positions. Since CLS reconstruction is our final choice, the Patch reconstruction loss and initial design study are detailed in Appendix A.2.

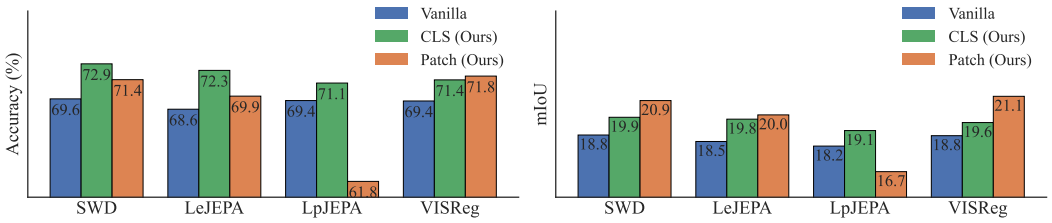

Figure 2: Effect of two-level token reconstruction on Top-1 KNN accuracy and linear probe segmentation. Both results indicate that both token-level reconstruction approaches significantly increase the model performance on both tasks. However, LpJEPA (Kuang et al., 2026) does not benefit more from patch token reconstruction, possibly due to its sparse nature.

Training objective. Using the shared-encoder projections in Eq. 2, with $\mathsf{D}(z,u) = \|z-u\|_2^2/d$ and $\bar{z} = G^{-1}\sum_i z_i$, global invariance is

$$\mathcal{L}_{\text{global}} = \frac{1}{G}\sum_{i=1}^{G} \mathsf{D}(z_i, \bar{z}). \tag{12}$$

The contextual loss for u_{ik} and v_{ij} is

$$\begin{aligned} \widehat{\mathcal{L}}_{\text{ctx}} &= \frac{1}{G}\sum_{i=1}^{G} \ell_i, \\ \ell_i &= \frac{\sum_{k=1}^{K} \mathsf{D}(\text{sg}(z_i), u_{ik}) + \sum_{j=1}^{a} \mathsf{D}(z_i, v_{ij})}{a+K}. \end{aligned} \tag{13}$$

Here $\text{sg}(\cdot)$ denotes stop-gradient. The encoder objective for the CLS approach is

$$\mathcal{L} = (1-\lambda)\big(\mathcal{L}_{\text{global}} + \widehat{\mathcal{L}}_{\text{ctx}}\big) + \lambda\mathcal{R}, \tag{14}$$

Losses are averaged over images and $\mathcal{R}$ is applied to global, local, and dropped CLS projections. The standard configuration uses $K = 1$.

Section 3 characterizes the prediction requirements and shared-backbone gradients of this objective. Although contextual learning gains are possible without SG, we retain it in the final method for its stronger empirical performance (Section A.1.3).

4.2 Applicability Verification

This paper aims to design a general method that is applicable to heuristic-free JEPA methods, so we further analyze the effect of the proposed candidates on three additional methods: SIGReg (Balestriero & LeCun, 2025), VISReg (Wu et al., 2026), and LpJEPA (Kuang et al., 2026). The uniqueness of each method is described in Section 4.1.

We use the training and evaluation protocol in Appendix A.2. To obtain an insightful pattern, we sweep the hyperparameter in a small range to get relatively good results for each method. Vanilla and Patch share the same best training settings: {learning rate, λ} of {9e-4, 0.9} for SWD, VISReg and LpJEPA and {9e-4, 0.05} for SIGReg, with a dropping ratio of $[0.3, 0.7]$. For CLS, the unique changes for each method are: {2e-3, 0.9} for SWD, {2e-3, 0.1} for SIGReg, {1.5e-3, 0.9} for LpJEPA, and {9e-4, 0.9} with a dropping ratio of $[0.5, 0.9]$ for VISReg. Figure 2 shows the KNN accuracy and mIoU of the three settings across the four methods. There are two interesting observations from the results: 1) Both CLS and Patch approaches benefit the methods that create dense embeddings, where the asymmetric patch token reconstruction consistently outperforms the CLS token reconstruction on segmentation. 2) The patch token reconstruction hurts the performance of LpJEPA, which creates sparse embeddings, whereas the CLS token reconstruction increases the LpJEPA performance on both tasks. We hypothesize that the degradation arises from an interaction between patch-level reconstruction and the sparsity encouraged by LpJEPA. CLS-level reconstruction may be less sensitive to this interaction because it operates on an aggregated representation. Since the goal of the proposed method is to be broadly applicable, the CLS token approach is our final choice for AhJEPA design. Our study in Appendix A.2 also shows that improvements are not because more views are involved.

Q. Is there an approach that is applicable to heuristic-free JEPA methods?
A. Yes, randomly dropping patch tokens and reconstructing on the CLS token.

5 EXPERIMENT

This section comprehensively evaluates the AhJEPA method on 22 datasets, covering the linear probe of in-domain and OOD classification, image segmentation, depth estimation, and latent world model planning.

5.1 PRETRAINING AND DOWNSTREAM EVALUATION

Training settings. Due to the limited computing resources, we only run ViT-B/16 (Dosovitskiy et al., 2021) pre-training on ImageNet (Deng et al., 2009) with VISReg (Wu et al., 2026) and SWD (Bonneel et al., 2015) because they are similar enough to use the same hyperparameter settings, and the result on VISReg gives us a direct comparison of the efficacy of AhJEPA. The invariance loss is calculated with 2 global views and 8 local views, and the patch token masking and reconstruction are applied to each global view. The pretraining step is finished with 8 H100s. More details are in Appendix A.3.

Datasets. Depending on the usage, our datasets can be categorized into pretraining, downstream evaluation, and latent world model planning. Dataset details are in Appendix A.4.

Pretraining: Following the previous studies, we pretrain the models on ImageNet1K (Deng et al., 2009) and use the validation set to monitor the training health.

Downstream tasks: For in-domain linear probe evaluation, we use FGVC-aircraft (Maji et al., 2013), Stanford cars (Krause et al., 2013), Cifar10 & Cifar100 (Krizhevsky et al., 2009), Oxford 102 flowers (Nilsback & Zisserman, 2008), Food 101 (Bossard et al., 2014), Oxford-IIIT Pet (Parkhi et al., 2012), and ImageNet1K. For OOD linear probe evaluation, we use Describable Textures Dataset (DTD) (Cimpoi et al., 2014), Galaxy10 (Leung, 2025), ChestXRay (Wang et al., 2017), Aerial Image Dataset (AID) (Xia et al., 2017), RetinaMNIST (Liu et al., 2022), and OrganAMNIST (Bilic et al., 2023). These two sets of datasets comprehensively reflect the representation separability. For dense segmentation, we follow DINOv2 (Oquab et al., 2024), applying linear segmentation and linear depth estimation on ADE20K (Zhou et al., 2017), Pascal VOC (Everingham et al.), NYU-Depth v2 (Nathan Silberman & Fergus, 2012) and KITTI (Eigen et al., 2014).

Latent world model planning: We follow LeWorldModel (Maes et al., 2026), using *PushT* (Zhou et al., 2025), *Reacher* (Tassa et al., 2018), *OGBench-Cube* (Park et al., 2025), and *TwoRoom* (Sobal et al., 2025). All datasets are offline trajectories stored as HDF5 files, with RGB observations rendered at 224×224 and a frameskip of 5. We use a 90/10 train/validation split (seed 3072), and each training sample is a 4-frame window (3 context frames and 1 prediction target).

Evaluation settings. For all linear classification tasks, we run full-shot on the training split for 10 epochs and calculate the accuracy on the test split. Despite previous works reporting the results of 100 epochs, we find the rank order barely changes, and it takes much longer time. For linear segmentation and depth estimation tasks, we use the output of the last layer and the other parts follow DINOv2. We report mIoU, RMSE and δ_1 for segmentation and depth estimation. To evaluate goal-conditioned control, we measure the success rate within a fixed step budget. The agent's action sequences are optimized using the Cross-Entropy Method (CEM) within a receding-horizon Model Predictive Control (MPC) framework. The planning horizon is set to 5 steps, equivalent to 25 environment time steps due to frame skipping.

In-domain classification accuracy. To ensure fairness, we mostly compare to the other methods at ViT-B/16 scale. The methods are categorized into *w/ heuristics* for training stability: (Baevski et al., 2022; Assran et al., 2023; Chen et al., 2021; Caron et al., 2021; Zhou et al., 2021), and *w/o heuristics* for training stability: (He et al., 2022; Wu et al., 2026) and ours. The intra-category comparison shows that AhJEPA achieves the best accuracy on six out of eight datasets, and the average accuracy has a non-trivial, *i.e.*, 1%, increase. The inter-category comparison shows that AhJEPA achieves

Table 1: In-domain linear probe classification results. Methods are pre-trained on Inet1K and a linear layer is trained for 10 epochs on each dataset. The accuracy on Inet1K is directly pulled from the corresponding work. Our AhJEPA framework improves the average accuracy of VISReg by 1%.

Methods	ViT	Epochs	Aircraft	Cars	Cifar10	Cifar100	Flowers	Food	Pets	Inet1K*	Avg.
w/ heuristics											
data2vec	L/14	1600	43.9	38.7	96.9	83.7	81.4	79.6	83.0	77.3	73.1
I-JEPA	H/14	300	55.4	59.2	97.2	85.5	86.8	83.3	92.8	79.3	79.9
MoCoV3	B/16	300	57.9	67.5	96.9	85.2	91.5	81.8	89.8	76.7	80.9
DINO	B/16	400	63.6	73.9	96.5	85.0	94.6	83.1	93.6	78.2	<u>83.6</u>
iBOT	B/16	400	63.5	73.8	97.1	85.9	93.7	84.2	93.6	79.5	**83.9**
w/o heuristics											
MAE	L/16	1600	61.9	61.5	93.3	78.0	85.4	78.6	91.3	75.1	78.1
VISReg	B/16	400	57.1	64.8	94.6	78.8	90.4	82.9	88.3	75.7	79.1
Ours (*VIS*)	B/16	400	56.8	65.8	95.9	83.4	91.0	83.2	88.7	76.3	80.1
Ours (*SWD*)	B/16	400	57.2	66.4	96.3	82.4	90.6	83.3	89.4	76.4	80.3

Table 2: OOD linear probe classification results. It covers texture (Cimpoi et al., 2014), space (Leung, 2025), aerial (Xia et al., 2017), and medical (Wang et al., 2017; Liu et al., 2022; Bilic et al., 2023) domains. AhJEPA helps heuristic-free methods achieve the best accuracy and improves VISReg's accuracy by 0.29%.

Methods	ViT	DTD	Galaxy10	AID	ChestXRay	Retina.	OrganA.	Avg.
w/ heuristics								
data2vec	L/16	69.68	65.73	85.98	22.45	63.75	89.04	66.10
I-JEPA	H/14	69.89	71.31	88.68	23.46	**65.75**	92.19	68.55
MoCoV3	B/16	73.72	73.06	90.20	23.89	<u>64.50</u>	91.37	69.46
DINO	B/16	74.26	72.77	<u>91.52</u>	24.63	62.50	91.70	69.56
iBOT	B/16	74.10	71.65	**91.73**	24.80	63.75	91.84	69.64
w/o heuristics								
MAE	L/16	72.77	71.98	86.42	22.87	63.00	90.06	67.85
VISReg	B/16	75.69	<u>74.01</u>	90.91	<u>24.88</u>	62.25	**93.40**	70.19
Ours (*SWD*)	B/16	<u>75.74</u>	73.79	90.80	24.45	64.00	92.99	<u>70.30</u>
Ours (*VIS*)	B/16	**75.90**	**74.35**	90.94	**25.15**	63.50	<u>93.05</u>	**70.48**

on-par or better performance than data2vec, I-JEPA, and MoCoV3, but the accuracy gap to iBOT still remains large, around 4%. Further details are in Table 1.

OOD classification accuracy. Performing well on OOD datasets is one of the major contributions of regularization-based methods (Balestriero & LeCun, 2025; Wu et al., 2026). But adding a reconstruction loss could force the model to learn in-domain details and hurt the OOD generalizability. To this end, we follow VISReg (Wu et al., 2026)'s OOD test to verify its effect. Table 2 indicates that AhJEPA is also helpful for OOD tasks, which increases the average accuracy on OOD test sets by 0.29%. This implies that AhJEPA is not just forcing the model to learn the details of the training data, but also helping it learn a general latent representation.

Segmentation and depth estimation. This directly reflects how well a representation preserves semantic and spatial information. The observation is that AhJEPA improves the performance on both tasks but still has a gap to iBOT on segmentation. The left block in Table 3 shows that AhJEPA improves the performance of VISReg by 0.89% on ADE20K and 1.94% on VOC, reflecting the efficacy of the proposed method. Compared to the other methods, despite the large gap to iBOT, the result is still very competitive on the segmentation task. The right block in Table 3 indicates that regularization-based JEPA methods are strong in depth estimation, and AhJEPA helps them perform better on NYU Depthv2, a harder dataset. Considering AhJEPA's one-branch design, it is less computationally expensive to scale up, which potentially leads to a better result.

Latent world model. Another very popular and challenging task is latent world model planning. We directly apply AhJEPA on all frames and add the reconstruction loss to the prediction loss. We find that it improves the success rate in OGBench-Cube and Two-Room by 4% and 11%, respectively.

Table 3: Linear segmentation and depth estimation. We report mIoU, RMSE, and δ_1 to measure performance. AhJEPA achieves competitive results compared to the other methods for both SWD and VISReg.

	ADE20K	VOC	NYU Depth v2		KITTI Eigen	
Method	mIoU↑	mIoU↑	RMSE ↓	δ_1 ↑	RMSE ↓	δ_1 ↑
w/ heuristics						
DINO	30.47	64.94	0.707	0.708	5.463	0.781
MoCoV3	31.69	<u>71.34</u>	0.704	0.714	5.420	0.755
data2vec	21.99	58.83	0.779	0.655	5.730	0.794
iBOT	**35.48**	**73.98**	**0.644**	<u>0.752</u>	5.030	<u>0.806</u>
w/o heuristics						
MAE	23.60	54.30	0.778	0.674	5.031	0.802
VISReg	31.36	66.15	0.688	0.725	**4.803**	**0.819**
Ours (*SWD*)	<u>32.49</u>	68.32	<u>0.647</u>	**0.753**	5.411	0.766
Ours (*VIS*)	32.25	68.09	0.663	0.744	<u>4.820</u>	0.798

Table 4: Latent world model planning. AhJEPA improves LeWM (Maes et al., 2026)'s performance to state-of-the-art on two datasets.

Method	PushT	Reacher	OGBench-Cube	Two-Room
Temporal (Wang et al., 2026)	91	-	-	-
FICReg (Seo et al.)	94	81	72	87
SD-JEPA (Thil et al., 2026)	97	**88**	72	90
LeWM (Maes et al., 2026)	<u>96</u>	<u>86</u>	74	87
AhJEPA-LeWM	<u>96</u>	84	**78**	**98**

Meanwhile, it maintains PushT performance, but affects Reacher by 2%. Considering that AhJEPA helps LeWM achieve SOTA performance on OGBench-Cube and Two-Room, the 2% decrease on Reacher is acceptable. Further details are in Table 4.

6 Conclusions

We mathematically characterize the prediction requirements of contextual learning and derive their implementation through a shared backbone without a separately parameterized teacher or EMA updates. Our analysis identifies the target-adaptation pathway removed by SG, while exploratory experiments show that fine-grained transfer gains are possible without it. We propose AhJEPA, a method that forces the model to learn granular details and applies to heuristic-free JEPA methods. It maintains the one-branch training paradigm and reduces the computational overhead compared to the popular MIM methods. Experimental results indicate that AhJEPA helps learn better representations, as tested on in-domain and OOD classification, segmentation, depth estimation, and latent world model planning.

Despite its effectiveness, the performance gap on segmentation and in-domain classification is still large compared to the SOTA model. We have tried to bring in registers (Darcet et al., 2024) and dedicated layer norms(Marouani et al., 2026), but both involve a trade-off between downstream tasks. The current AhJEPA approach is designed as a broadly applicable method with a single-branch training paradigm rather than a highly optimized, state-of-the-art implementation tailored to a single setting. To achieve the best performance, a proper optimization design of patch token reconstruction for methods that create dense representations is needed for future work. In addition, a more effective method for improving fine-grained performance without using stop-gradient is our next goal, given the theoretical analysis and the good results of our initial attempt.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, proof writing, method implementation, hyperparameter sweeps, result collection, and literature searches to assess novelty. The core method design, experimental design, and analysis of the results presented in the paper were conducted entirely by the authors. All AI-assisted content was reviewed by the authors. We take full responsibility for the final manuscript, including its mathematical claims, proofs, experimental results, and code.

REPRODUCIBILITY STATEMENT

The heuristic-free JEPA methods are all publicly available, and we have released all the details of our method, including architecture, loss function, hyperparameters, datasets, and training/test configurations.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022.

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

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *ICLR*, 2022.

Patrick Bilic, Patrick Christ, Eugene Vorontsov, and et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 2023.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pp. 9640–9649, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

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

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *3dRR*, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yilun Kuang, Yash Dagade, Tim GJ Rudner, Randall Balestriero, and Yann LeCun. Rectified lpjepa: Joint-embedding predictive architectures with sparse and maximum-entropy representations. *arXiv preprint arXiv:2602.01456*, 2026.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 2022.

Henry Leung. Galaxy10 DECaLS Dataset. `https://astronn.readthedocs.io/en/latest/galaxy10.html`, 2025. Accessed: 2026-01-11.

Rui Liu, Xinhui Wang, Qiang Wu, and et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 2022.

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

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy Vo. Revisiting [cls] and patch token interaction in vision transformers. In *International Conference on Learning Representations*, volume 2026, pp. 22245–22263, 2026.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *T-MLR*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *ICLR*, 2025.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

Sungwon Seo, Jean Seong Bjorn Choe, and Jong-Kook Kim. Ficreg: Forward-inverse consistency regularization for latent world models. In *Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning*.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Stress-testing offline reward-free reinforcement learning: A case for planning with latent dynamics models. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Subspace-decomposed jepas: Disentangling progression and content in latent world models. *arXiv preprint arXiv:2605.31111*, 2026.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

Ying Wang, Oumayma Bounou, Gaoyue Zhou, Randall Balestriero, Tim G. J. Rudner, Yann LeCun, and Mengye Ren. Temporal straightening for latent planning. *ICML*, 2026.

Haiyu Wu, Randall Balestriero, and Morgan Levine. Visreg: Variance-invariance-sketching regularization for jepa training. *arXiv preprint arXiv:2606.02572*, 2026.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.*, 2017.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

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

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2021.

A Appendix

A.1 Proofs and supplementary contextual analysis

This section provides the proofs of the results in Section 3 and supplementary results for fixed output moments and complementary views.

A.1.1 Proofs

Proof of Lemma 1.

Proof. Write $m = \mathbb{E}[z \mid V]$. Since m is determined by V,

$$\mathbb{E}[z - m \mid V] = \mathbb{E}[z \mid V] - m = 0.$$

Both m and q are functions of V. Conditioning the cross term on V therefore gives

$$\begin{aligned}\mathbb{E}[(z-m)^\top(m-q)] &= \mathbb{E}\big[\mathbb{E}[(z-m)^\top(m-q) \mid V]\big] \\ &= \mathbb{E}\big[\mathbb{E}[z-m \mid V]^\top(m-q)\big] = 0.\end{aligned}$$

Using $z-q=(z-m)+(m-q)$, expand

$$\begin{aligned}\mathbb{E}\|z-q\|_2^2 &= \mathbb{E}\|z-m\|_2^2 + \mathbb{E}\|m-q\|_2^2 \\ &\quad + 2\mathbb{E}[(z-m)^\top(m-q)] \\ &= \mathbb{E}\|z-m\|_2^2 + \mathbb{E}\|m-q\|_2^2.\end{aligned}$$

Dividing by d proves the identity. □

Proof of Theorem 1.

Proof. Let $m_L=\mathbb{E}[z \mid V_L]$ and $\mu=\mathbb{E}[z]$. The tower property gives $\mathbb{E}[m_L]=\mu$ and

$$\begin{aligned}\mathbb{E}[(z-m_L)^\top(m_L-\mu)] &= \mathbb{E}\big[\mathbb{E}[z-m_L \mid V_L]^\top(m_L-\mu)\big] \\ &= 0.\end{aligned}$$

Expanding $z-\mu=(z-m_L)+(m_L-\mu)$ gives

$$\begin{aligned}\sigma_z^2 &= \mathbb{E}\|z-m_L\|_2^2 + \mathbb{E}\|m_L-\mu\|_2^2 \\ &= \mathbb{E}\|z-m_L\|_2^2 + \sigma_z^2\eta_L(z).\end{aligned}$$

Nonnegativity of both terms establishes $0\le\eta_L(z)\le 1$. Lemma 1, with $q=v$, yields

$$\begin{aligned}\mathcal{L}_L &= \frac{\sigma_z^2}{d}(1-\eta_L(z)) + \frac{1}{d}\mathbb{E}\|v-m_L\|_2^2 \\ &\ge \frac{\sigma_z^2}{d}(1-\eta_L(z)).\end{aligned}$$

Under identity covariance, $\sigma_z^2=\mathrm{tr}(I_d)=d$. Rearranging and using $\eta_L(z)\ge 0$ proves the final inequality. □

Proof of Proposition 1.

Proof. For a single dropped-view pair, the differential is

$$\begin{aligned}\mathrm{d}\,\mathsf{D}(z,u) &= \frac{2}{d}(z-u)^\top(\mathrm{d}z-\mathrm{d}u) \\ &= \frac{2}{d}(z-u)^\top(J_z-J_u)\,\mathrm{d}\theta.\end{aligned}$$

Consequently,

$$\nabla_\theta \mathsf{D}(z,u) = \frac{2}{d}\left[J_u^\top(u-z)+J_z^\top(z-u)\right].$$

Replacing u by v gives the local-view gradient. Summing over the K dropped and a local pairs and dividing by $a+K$ proves Eq. 7. To identify the parent contribution at the output level, hold the child outputs fixed and define

$$c_i = \frac{\sum_k u_{ik} + \sum_j v_{ij}}{a+K}.$$

Then

$$\begin{aligned}\frac{\partial \ell_i^{\mathrm{full}}}{\partial z_i} &= \frac{2}{d(a+K)}\left[\sum_k (z_i-u_{ik}) + \sum_j (z_i-v_{ij})\right] \\ &= \frac{2}{d}(z_i-c_i).\end{aligned}$$

Thus, at the output level, gradient descent moves the parent toward the mean child prediction. □

Proof of Proposition 2.

Proof. Let $r = z - m_\theta(V)$ and $J_m = \partial m_\theta(V)/\partial\theta$. The distribution of observations is independent of θ, so differentiation under the expectation gives

$$\nabla_\theta \varepsilon_V = 2\mathbb{E}[(J_z - J_m)^\top r].$$

Since J_m is a function of V and $\mathbb{E}[r \mid V] = 0$,

$$\mathbb{E}[J_m^\top r] = \mathbb{E}[J_m^\top \mathbb{E}[r \mid V]] = 0.$$

Consequently, $\nabla_\theta \varepsilon_V = 2\mathbb{E}[J_z^\top r]$. Substituting $z - u = r + (m_\theta(V) - u)$ yields

$$\begin{aligned}\frac{2}{d}\mathbb{E}[J_z^\top (z-u)] &= \frac{2}{d}\mathbb{E}[J_z^\top r] + \frac{2}{d}\mathbb{E}[J_z^\top (m_\theta(V) - u)] \\ &= \frac{1}{d}\nabla_\theta \varepsilon_V + \frac{2}{d}\mathbb{E}[J_z^\top (m_\theta(V) - u)].\end{aligned}$$

Finally, $\mathrm{sg}(z_i) = z_i$ in the forward pass, but its backward derivative is zero. In Eq. 7, this deletes exactly the $J_{z_i}^\top (z_i - u_{ik})$ terms. Averaging over parents gives Eq. 10; its coefficient in the total gradient is $1 - \lambda$. □

Prediction requirements of CLS reconstruction. To understand what contextual reconstruction adds to the base JEPA objective, we first separate output statistics from cross-view relationships. The following proposition shows that, once output moments are fixed, reconstruction error measures agreement between complete and restricted observations.

Proposition 3 (Alignment under fixed output moments)**.** *Under the idealized conditions* $\mathbb{E}[z] = \mathbb{E}[q] = 0$ *and* $\mathrm{Cov}(z) = \mathrm{Cov}(q) = I_d$*,*

$$\mathbb{E}\,\mathsf{D}(z, q) = 2 - \frac{2}{d}\,\mathrm{tr}\,\mathbb{E}[zq^\top]. \tag{15}$$

Thus, on this constrained set, minimizing alignment error is equivalent to maximizing aggregate cross-view correlation, including when both outputs share parameters.

This motivates adding contextual reconstruction to distributionally regularized JEPA methods: the regularizer controls output statistics, while reconstruction constrains the relationships encoded across views. The assumed moments follow from ideal matching to the standard Gaussian reference of SIGReg, SWD, and VISReg; they are not exact minibatch constraints or assumptions about iBOT and patch features. The variance analysis in Section 3 does not require Gaussianity and also covers embeddings trained with LpJEPA.

Proof of Proposition 3.

Proof. The assumed moments imply

$$\mathbb{E}\|z\|_2^2 = \mathrm{tr}\,\mathrm{Cov}(z) + \|\mathbb{E}[z]\|_2^2 = d,$$

and likewise $\mathbb{E}\|q\|_2^2 = d$. Since $z^\top q = \mathrm{tr}(zq^\top)$,

$$\begin{aligned}\mathbb{E}\,\mathsf{D}(z, q) &= \frac{1}{d}\left(\mathbb{E}\|z\|_2^2 + \mathbb{E}\|q\|_2^2 - 2\mathbb{E}[z^\top q]\right) \\ &= 2 - \frac{2}{d}\,\mathrm{tr}\,\mathbb{E}[zq^\top].\end{aligned}$$

The first term is constant, establishing the equivalence. The calculation does not require separate parameterizations. □

A.1.2 COMPLEMENTARY PREDICTION REQUIREMENTS

Corollary 1 (Complementary prediction requirements)**.** *Sample a parent uniformly and one of its dropped or local views uniformly within each type. Let $\mathcal{L}_D, \mathcal{L}_L$ be the corresponding errors, and define $\eta_D(z)$ as in Eq. 4 with a dropped observation. For $a, K > 0$,*

$$\begin{aligned}\mathcal{L}_{\mathrm{ctx}} := \mathbb{E}\widehat{\mathcal{L}}_{\mathrm{ctx}} &= \frac{K\mathcal{L}_D + a\mathcal{L}_L}{a+K} \\ &\geq \frac{\sigma_z^2}{d}\frac{K(1-\eta_D(z)) + a(1-\eta_L(z))}{a+K}.\end{aligned} \tag{16}$$

In particular, under $\mathrm{Cov}(z) = I_d$,

$$\eta_L(z) \geq \max\left\{0, 1 - \frac{a+K}{a}\mathcal{L}_{\mathrm{ctx}}\right\}. \tag{17}$$

Proof. Taking expectations in Eq. 13 gives the weighted average. Applying Theorem 1 to each observation type gives

$$\mathcal{L}_D \geq \frac{\sigma_z^2}{d}(1-\eta_D(z)), \qquad \mathcal{L}_L \geq \frac{\sigma_z^2}{d}(1-\eta_L(z)).$$

Multiply these inequalities by $K/(a+K)$ and $a/(a+K)$, respectively, and add them to obtain Eq. 16. Furthermore, $\mathcal{L}_D \geq 0$ implies

$$\mathcal{L}_L = \frac{(a+K)\mathcal{L}_{\mathrm{ctx}} - K\mathcal{L}_D}{a} \leq \frac{a+K}{a}\mathcal{L}_{\mathrm{ctx}}.$$

Under identity covariance, substitution into the theorem gives

$$\eta_L(z) \geq 1 - \mathcal{L}_L \geq 1 - \frac{a+K}{a}\mathcal{L}_{\mathrm{ctx}}.$$

Combining this with $\eta_L(z) \geq 0$ proves the result. □

Together, these bounds characterize the relationships encouraged by the CLS objective: full-view information must be predictable both under token dropping and from local evidence. Low error for one observation type does not remove the other's prediction requirement. Sampling both types from a common parent connects these requirements to the same image content. This provides a mechanism for learning relationships between local evidence and broader visual structure, whose benefits for fine-grained tasks are evaluated in Section 5.

A.1.3 AN EXPLORATORY ALTERNATIVE TO STOP-GRADIENT

The decomposition in Section 3.3 suggests regulating joint prediction and target adaptation as an alternative to deleting the target gradient. We explore this possibility using two complementary retained subsets of each global view.

Complementary-subset objective. With $K = 2$ and $\bar{u}_i = K^{-1}\sum_k u_{ik}$, define

$$\mathcal{L}_{D,i}^{(\beta)} = \mathsf{D}(z_i, \bar{u}_i) + \frac{\beta}{K}\sum_k \mathsf{D}(u_{ik}, \bar{u}_i). \tag{18}$$

The first term aligns the mean subset prediction with the parent, and the second controls agreement between subset predictions. Both sides remain differentiable. We replace the dropped-view sum in ℓ_i^{full} by $K\mathcal{L}_{D,i}^{(\beta)}$ to preserve its weighting relative to local views.

This modification separates two components already present in pairwise reconstruction. Since $\sum_k(u_{ik} - \bar{u}_i) = 0$, expanding $z_i - u_{ik} = (z_i - \bar{u}_i) - (u_{ik} - \bar{u}_i)$ gives

$$\begin{aligned}\frac{1}{K}\sum_k \|z_i - u_{ik}\|^2 = \|z_i - \bar{u}_i\|^2 &+ \frac{1}{K}\sum_k \|u_{ik} - \bar{u}_i\|^2 \\ &- \frac{2}{K}(z_i - \bar{u}_i)^\top \underbrace{\sum_k (u_{ik} - \bar{u}_i)}_{=0}.\end{aligned}$$

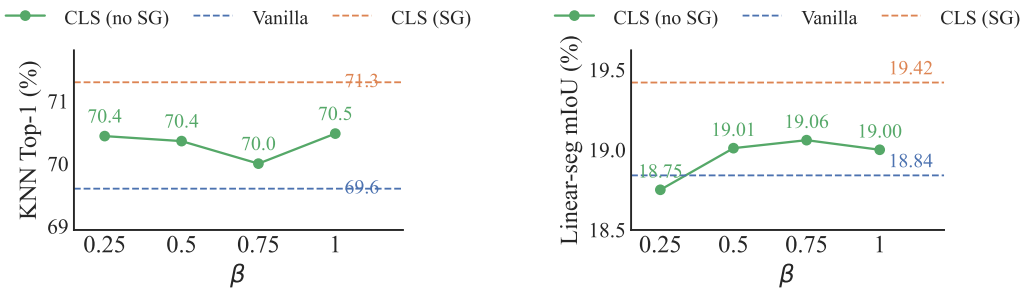

Figure 3: Contextual learning without SG. The solid curve shows complementary-subset configurations with $2g + 6l + 4d$. Dashed references show vanilla training with $2g + 8l$ and AhJEPA with SG using $2g + 8l + 2d$. Here g, l, and d count global, local, and dropped views. The configurations use different view allocations.

Dividing by d shows that $\beta = 1$ recovers ordinary per-subset alignment. Varying β therefore controls subset agreement separately from mean prediction, without removing the target gradient. In particular, $\beta < 1$ relaxes agreement between individual subsets while retaining alignment of their mean with the global embedding. This gives a concrete alternative to explore, rather than a guarantee that joint adaptation will retain useful details.

Exploratory results. Using the ImageNet-100 and ADE20K protocol described in Appendix A.2, Figure 3 shows that this configuration can improve fine-grained transfer without SG. At $\beta = 1$, kNN accuracy increases from 69.6% to 70.5% and ADE20K mIoU from 18.84% to 19.00% over vanilla training. The comparison establishes feasibility with the reported view allocations, rather than isolating the effect of SG alone. AhJEPA with SG achieves the best results in this comparison (71.30% kNN and 19.42% mIoU). We therefore retain SG in our final method, while leaving more effective regularization of joint adaptation to future work.

A.2 One-branch reconstruction design study

iBOT (Zhou et al., 2021) improves DINO (Caron et al., 2021) method on fine-grained tasks by applying masked token reconstruction. Due to its teacher-student architecture with a bag of heuristic tricks, it is hard to apply this method to other domains. SIGReg (Balestriero & LeCun, 2025) and its variations (Wu et al., 2026; Kuang et al., 2026) provide a one-branch and heuristic-free approach, but the vanilla loss does not force the model to learn from context, resulting in weak segmentation performance (Wu et al., 2026).

For the Patch candidate, let $\mathcal{B}$ index clean–masked global-view pairs in a minibatch and $\mathcal{M}_b$ denote the masked positions of $\widehat{x}_b$. Writing $h_{\theta,r}$ for the backbone feature at position r, the reconstruction loss is

$$\mathcal{L}_{\text{mim}}^{\text{Patch}} = \frac{\sum_{b\in\mathcal{B}} \sum_{r\in\mathcal{M}_b} \|p_{br} - \text{sg}(t_{br})\|_2^2}{d_p \sum_{b\in\mathcal{B}} |\mathcal{M}_b|}, \tag{19}$$

where $p_{br} = P_\theta(h_{\theta,r}(\widehat{x}_b))$ is the masked-patch MLP prediction, $t_{br} = \text{LN}(h_{\theta,r}(x_b))$ is the clean target with parameter-free LayerNorm, and d_p is their feature dimension. The loss averages over all masked positions in the minibatch. The CLS objective used by AhJEPA is defined in Section 4.

Experimental settings. We evaluate online linear-probe accuracy, kNN accuracy, and segmentation mIoU. Unless otherwise specified, pretraining uses ImageNet-100 (Deng et al., 2009), a ViT-S/16 backbone (Dosovitskiy et al., 2021), and 100 epochs with two 224×224 global views and eight 96×96 local views. Patch-masking candidates use iBOT-style masking with the ratio sampled uniformly from $[0.3, 0.7]$. The CLS variant instead drops a uniformly sampled set of tokens, with the drop ratio drawn from the same interval. For SWD, the learning rate is 9×10^{-4} and $\lambda = 0.9$.

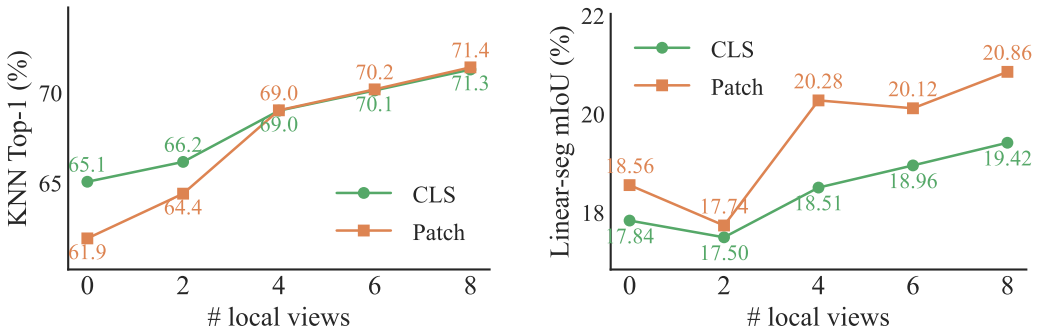

Figure 5: Effect of local views in two-level token reconstruction on KNN and linear segmentation. We observe that more local views increase the performance of both approaches, and patch token reconstruction benefits more from it.

For dense evaluation, we train a linear segmentation head for 40 epochs on ADE20K (Zhou et al., 2017) using the DINOv2 (Oquab et al., 2024) evaluation setup.

Figure 4 summarizes the initial online linear-probe results. Neither the shared patch MLP nor the separate EMA-updated MLP improves on the unmasked baseline. Disabling weight decay for the separate MLP improves accuracy, but introduces an additional optimizer-specific choice. We therefore investigate CLS reconstruction and asymmetric patch reconstruction as alternative designs. Both improve linear-probe accuracy and are retained for further evaluation. We then compare kNN accuracy and segmentation mIoU while varying the number of local pretraining views. Figure 5 shows that patch reconstruction is effective even without local views, and that adding local views improves both metrics for both candidates. Specifically, as the number of local views increases, the KNN results of both approaches converge to the same point, but the mIoU gap on segmentation remains large.

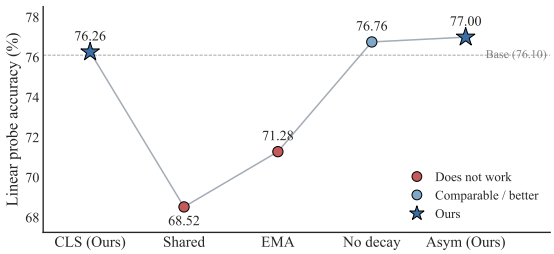

Figure 4: Road map of design trajectory. We evaluate five different reconstruction approaches by online linear probe accuracy. This shows that our asymmetric MLP patch token reconstruction design outperforms the others. Base is trained without masking.

Is the gain only from more views? Dropped views add encoder passes, so part of the improvement could come from the extra compute rather than from contextual reconstruction. We therefore compare AhJEPA with a vanilla baseline that uses the same number of views. The training settings are the same as previous experiments. As shown in Table 5, these two settings have comparable cost: AhJEPA needs 16% more FLOPs and 27% more peak memory per step. With this budget, AhJEPA improves kNN accuracy by 1.3 points and ADE20K mIoU by 0.32 points, whereas the extra local views do not close the gap. This indicates that the gain comes from the AhJEPA design rather than adding more views.

Table 5: Ablation on AhJEPA design. This table studies whether the improvements are from additional views or the method design. We report TFLOPs, memory cost, Top-1 KNN accuracy, and mIoU on ADE20K.

Setting	TFLOPs	Memory	KNN	mIoU
$2g + 10l$	6.31	8.37	70.00	19.10
$2g + 8l + 2d$	7.35	10.62	**71.30**	**19.42**

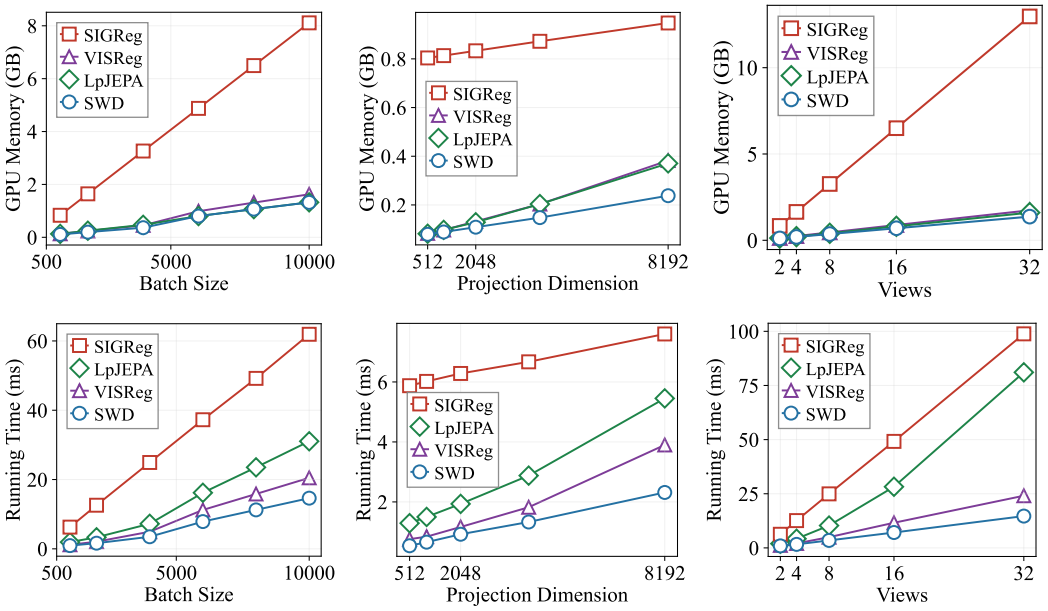

Figure 6: Scaling efficiency analysis for the regularization term in JEPA losses. The Optimal Transport methods (Bonneel et al., 2015; Kuang et al., 2026; Wu et al., 2026) are more efficient than the characteristic function method, *i.e.*, SIGReg (Balestriero & LeCun, 2025), on both memory and running time. The tests were run on a single RTX6K. We choose to use SWD in our investigation.

A.3 IMPLEMENTATION DETAILS

Table 6: Pretraining setup of drop-SG ViT-B on ImageNet-1K.

Config	Value
Architecture	
Backbone	ViT-B/16
Projection dim d_p	256
Data & augmentation (DINO-style multi-crop)	
Dataset	ImageNet-1K
Global crops	224^2, scale $[0.3, 1.0]$ of the image
Local crops	96^2, scale $[0.05, 0.3]$ of the parent global crop
Token-drop views	1 per global crop, drop ratio $\sim U(0.3, 0.7)$
Views per image	2 global + 8 local + 2 token-drop
Augmentations	h-flip, color jitter (p=0.8), grayscale (p=0.2), Gaussian blur (p=0.5), solarize (p=0.2)
Optimization	
Optimizer	AdamW (weight decay 5×10^{-2})
Learning rate	9×10^{-4}
LR schedule	5-epoch linear warmup, cosine to $\mathrm{lr}_{\max}/1000$
Gradient clipping	1.0
Epochs	400
Batch size	256

Table 7: LeWM world-model training setup.

Config	Value
Architecture	
Encoder	ViT-Tiny/14
Predictor	6-layer transformer (dropout 0.1)
SIGReg weight λ	0.1
Random projections M	1024
Data	
Environments	PushT, OGBench-Cube, Two-Room, Reacher
Observation size	224×224
Frame skip	5
Optimization	
Epochs	10 per environment
Batch size	128

Table 8: Evaluation setup. The encoder is frozen in all protocols.

Config	Downstream cls.	IN-1K linear probe	Dense prediction
Benchmark(s)	8 cls. + 6 OOD datasets	ImageNet-1K	ADE20K, Pascal VOC, KITTI, NYU Depth v2
Features	CLS of last 4 layers	CLS of last 4 layers	patch feat., last layer
Head	linear + SyncBN	multi-head linear + shared SyncBN (10 heads)	1×1 conv + SyncBN
Optimizer	SGD (mom. 0.9, no WD)	SGD (mom. 0.9, no WD)	AdamW
Learning rate	grid of 13 base LRs, linear scaling rule	10 LRs (one per head), linear scaling rule	2×10^{-3}
LR schedule	cosine annealing	per-step cosine annealing	poly decay (power 0.9)
Epochs	10	100	40 (seg) / 15 (depth)
Batch size	32	32 per GPU (256 eff., 8 GPUs)	16
Image size	224×224	train: RRC to 224×224; val: resize 256, crop 224	512×512
Metric	best test accuracy	best val. top-1 (50K)	mIoU (seg.) / RMSE (depth)

A.4 DATASETS

We evaluate our pre-trained models on a diverse set of 15 datasets spanning multiple domains, tasks, and difficulty levels. The following sections describe each dataset in detail.

ImageNet-1k (Deng et al., 2009) is a large-scale object recognition dataset containing 1,281,167 training images and 50,000 validation images across 1,000 classes, representing diverse natural objects, animals, and scenes from the natural world.

CIFAR-10 (Krizhevsky et al., 2009) is a 10-class object classification dataset with 50,000 training and 10,000 test images at 32×32 resolution, covering categories like airplanes, cars, birds, cats, and other common objects.

CIFAR-100 (Krizhevsky et al., 2009) is a fine-grained object classification dataset with 50,000 training and 10,000 test images at 32×32 resolution, containing 100 classes organized into 20 supercategories including various vehicles, animals, and household items.

Stanford Cars (Krause et al., 2013) is a fine-grained vehicle classification dataset with 8,144 training and 8,041 test images covering 196 car models from 98 manufacturers, spanning decades of automotive design from 1950 to 2012.

Galaxy10 (Leung, 2025) is an astronomical image classification dataset with 17,736 images classifying galaxies into 10 morphological categories (disturbed, merging, spiral, elliptical, etc.) from the DECaLS survey.

Food-101 (Bossard et al., 2014) is a food classification dataset with 75,750 training and 25,250 validation images covering 101 food categories including dishes like pizza, sushi, hamburger, and various international cuisines.

Oxford-IIIT Pets (Parkhi et al., 2012) is a pet breed classification dataset with 3,680 training and 3,669 test images covering 37 cat and dog breeds, requiring fine-grained distinction between similar-looking breeds.

NIH Chest X-ray (Wang et al., 2017) is a multi-label medical image classification dataset comprising 112,120 total images, where each chest radiograph may contain multiple pathology labels from 14 disease categories including pneumonia, cardiomegaly, and pleural effusion.

RetinaMNIST (Liu et al., 2022) is a medical image classification dataset with 1,080 training, 120 validation, and 400 test retinal fundus images classifying diabetic retinopathy into 5 severity grades.

OrganAMNIST (Bilic et al., 2023) is a medical image classification dataset with 34,581 training, 6,491 validation, and 17,778 test CT axial slices classifying 11 body organ types including liver, kidney, spleen, and heart.

Oxford Flowers 102 (Nilsback & Zisserman, 2008) is a fine-grained plant classification dataset with 1,020 training, 1,020 validation, and 6,149 test images covering 102 flower species with 40-258 images per class.

Describable Textures (DTD) (Cimpoi et al., 2014) is a texture classification dataset with 1,880 training, 1,880 validation, and 1,880 test images across 47 texture categories (e.g., braided, dotted, fibrous) following a 10-fold cross-validation protocol.

FGVC-Aircraft (Maji et al., 2013) is a fine-grained aircraft classification dataset with 6,667 trainval and 3,333 test images covering 100 aircraft variants from the FGVC-Aircraft 2013b benchmark.

AID (Xia et al., 2017) is a remote sensing scene classification dataset with 10,000 images across 30 aerial scene categories including airports, beaches, forests, and urban areas, using a 10%/90% train/test split for SSL evaluation.

ADE20K (Zhou et al., 2017) is a semantic segmentation dataset with 20,210 training and 2,000 validation images, containing pixel-level annotations for 150 semantic classes including objects, parts, and materials across indoor and outdoor scenes.

Pascal VOC 2012 (Everingham et al.) is a semantic segmentation dataset with 1,464 training and 1,449 validation images, containing pixel-level annotations for 20 object classes plus background, covering common objects such as people, animals, vehicles, and indoor items.

KITTI (Eigen et al., 2014) is a monocular depth estimation dataset captured from a vehicle-mounted camera and LiDAR in outdoor driving scenes, using the Eigen split with 23,488 training and 697 test images and sparse ground-truth depth up to 80 meters.

NYU Depth v2 (Nathan Silberman & Fergus, 2012) is a monocular depth estimation dataset of indoor scenes captured with a Kinect RGB-D sensor, containing 24,231 training and 654 test image-depth pairs from 464 scenes with dense ground-truth depth up to 10 meters.

PushT (Zhou et al., 2025) is a continuous 2D manipulation environment in which an agent must push a T-shaped block to match a target configuration through pushing interactions only; following the DINO-WM setup, the offline dataset contains 20,000 expert episodes with an average length of 196 steps.

OGBench-Cube (Park et al., 2025) is a continuous 3D robotic manipulation environment in which a robotic arm with an end-effector must pick up a cube and place it at a target location; we use the single-cube variant with 10,000 episodes of 200 steps each, collected with the heuristic policy provided in the benchmark.

Two-Room (Sobal et al., 2025) is a continuous 2D navigation environment consisting of two rooms separated by a wall with a single door, where an agent must navigate from a random start in one room to a target location in the other; the dataset contains 10,000 episodes with an average length of 92 steps, generated by a noisy heuristic policy.

Reacher (Tassa et al., 2018) is a continuous control environment from the DeepMind Control Suite in which a two-joint robotic arm must reach a target location in a 2D plane, with success defined by exact alignment of the arm joints with the target configuration; the dataset contains 10,000 episodes of 200 steps each, collected with a Soft Actor-Critic policy.

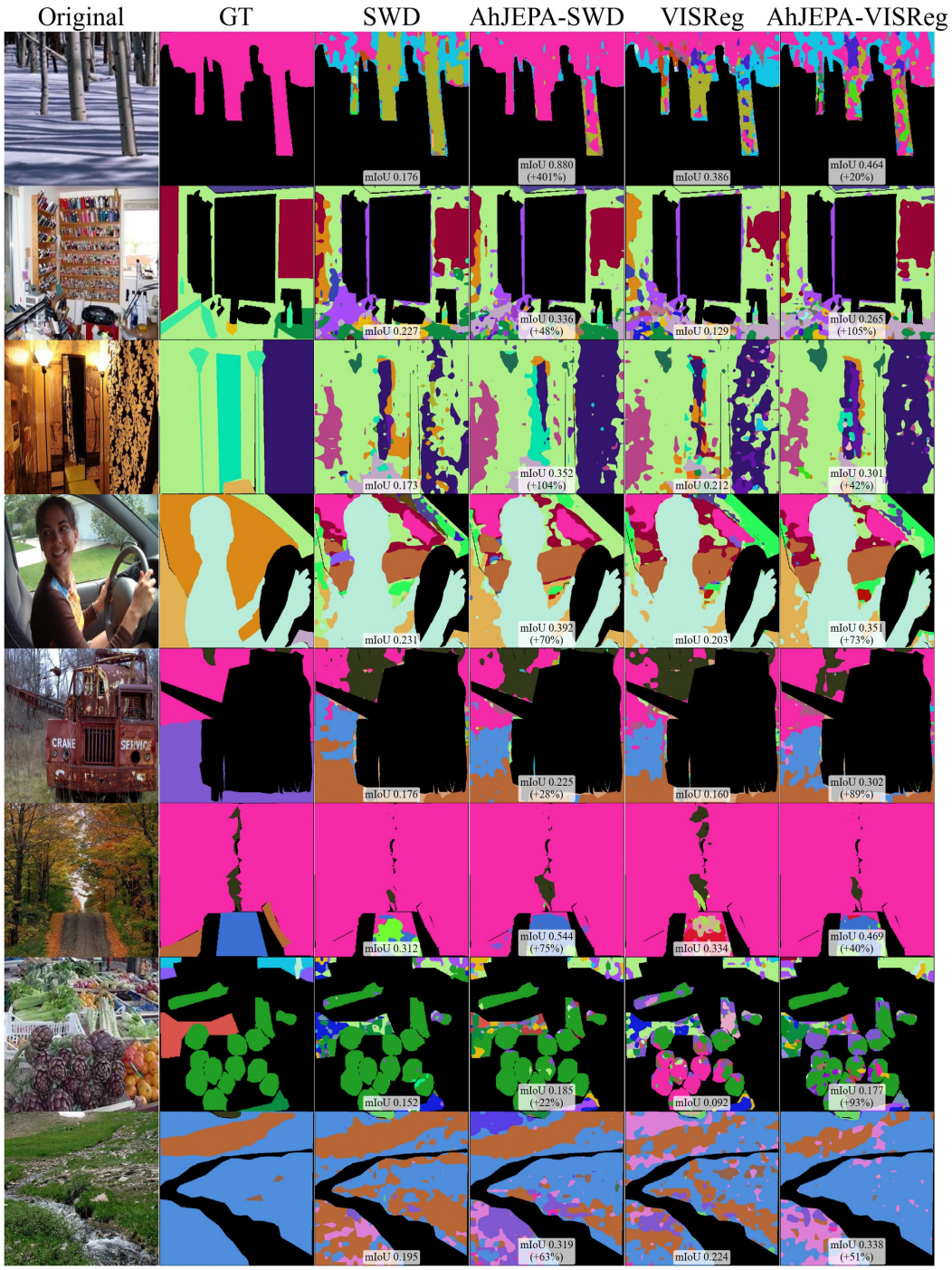

Figure 7: Examples of linear probe segmentation. Due to the low-quality visualization, we also report the improvements after applying AhJEPA.

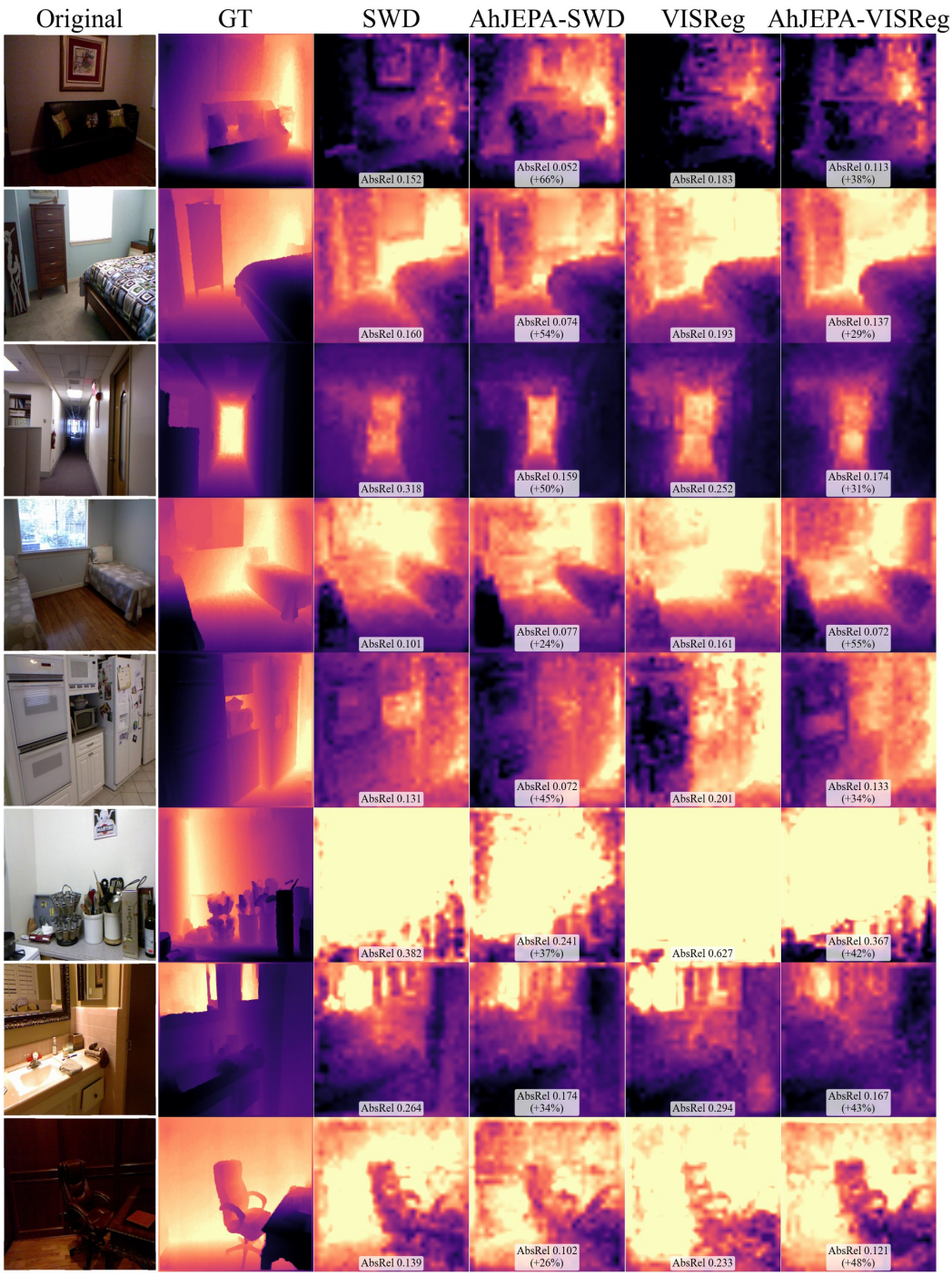

Figure 8: Examples of linear probe depth estimation. Due to the low-quality visualization, we also report the improvements after applying AhJEPA.