

Biomedical Self-Supervised Learning using Sequential and Action-Conditioned Prediction: A Proof of Concept in Histopathology

Linus Schilpp¹, Angelo Jovin Yamachui Sitcheu¹, Thomas Lautenschlager¹, Leonard Saur¹, Nils Friederich^{1,2*}, Ralf Mikut^{1*},
Stephan Allgeier^{1*}

¹Institute for Automation and Applied Informatics,
Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen

²Institute of Biological and Chemical Systems,
Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen

* Supervised equally

Abstract

Self-supervised foundation models for biomedical imaging are typically trained with non-sequential objectives that encourage invariance between augmented crops or predict missing content within a limited spatial resolution or range. This paradigm can be limiting for biomedical images, where diagnostically relevant structure often depends on relationships across widely different spatial extents and resolutions. We investigate sequential, action-conditioned prediction as an alternative self-supervised pretraining objective for biomedical imaging. As a proof of concept, we adapt seq-JEPA, originally developed for natural images, to 256×256 px histopathology images by extending its action space with zoom, and introduce seq-JEPA-dense, which predicts the next representation for

every step in the sequence. Under the same per-crop augmentation distribution and with one run per method, seq-JEPA-dense improved tissue classification accuracy in the aggregated linear-probe evaluation from 88.3% to 91.9%, narrowing the gap to SimCLR, which achieved 93.3%. Both sequential variants recover a held-out crop from the preceding crops and the relative action leading to it at roughly twice the chance rate within the same image, indicating that the prediction depends on the action; whether position, zoom, or both drive this remains to be isolated. These results provide proof-of-concept evidence that sequential, action-conditioned self-supervision can be adapted to histopathology by extending its action space with zoom, yielding representations that approach a contrastive baseline, a first step toward its investigation as a self-supervised approach for extended biomedical images such as whole-slide images.

1 Introduction

Self-supervised foundation models have been increasingly developed for biomedical images in recent years, enabling a path toward annotation-free learning from raw images, while generalizing to a variety of downstream tasks with only a small amount of annotated data. Such models can be trained by several paradigms: contrastive methods (e.g., SimCLR [1], with biomedical variants such as MICLe [2]), non-contrastive methods based on self-distillation (e.g., DINO [3], with biomedical variants such as RAD-DINO [4]), reconstruction-based masked image modeling (e.g., MAE [5], with biomedical variants such as CA-MAE [6]), and joint-embedding predictive architectures (JEPAs), which predict target representations in latent space rather than reconstructing pixels (e.g., I-JEPA [7], with biomedical variants such as RadJEPA [8]). Beyond purely image-based signals, self-supervised vision–language models exploit naturally paired image–text data as the supervisory signal (e.g., CLIP [9], with biomedical variants such as BiomedCLIP [10]).

Despite their diversity, image-based objectives share a common recipe. Training rests on single-step objectives that encourage invariance between augmented crops (e.g., SimCLR [1], DINO [3]), or predict missing content within a limited spatial resolution or range (e.g., MAE [5], I-JEPA [7]), in pixel or latent space,

using the remaining visible content. At inference, typically only the pretrained encoder is retained and applied to each input independently, followed by task-specific downstream modules such as classifiers, decoders, or aggregators.

Biomedical images often depict extended specimens whose informative content spans an area or a volume that a single model input typically cannot capture at both wide range and high resolution simultaneously within practical compute and memory budgets [11, 12]: an input at native resolution covers only a small region, while one that covers a larger extent of the specimen does so at reduced resolution [13]. In histopathology, foundation models typically operate at the small-input end of this trade-off, on tiles of 224×224 px that are far smaller than the field of view used in microscopic diagnosis, and have consequently been argued to capture fine-grained texture at the expense of mesoscale tissue organization [14]. We use the term *crop* for a subregion sampled at variable position and size as a model input, whereas a *tile* denotes one of the fixed-size, non-overlapping pieces into which a whole-slide image is partitioned on a regular grid to match the encoder’s input size. Single-step objectives can combine inputs from several points along this range–resolution trade-off, e.g., by training on crops of varying sizes [15], but they relate these inputs directly within one step, which limits how far apart in position and scale related inputs can be while remaining similar or directly predictable from one another. Both the training and the inference recipe described above are thus at odds with the structure of the data.

Furthermore, many self-supervised training objectives primarily optimize for invariance across augmented crops (e.g., SimCLR- and DINO-family models [3, 16, 17]) without explicitly modeling the transformations relating them. These relative relations are themselves diagnostically meaningful: whether two abnormal areas are part of one region or of two separate ones cannot be answered from either alone. Such invariance assumes that crops at different positions and scales depict the same object, as made explicit in local–global crop training, an assumption that has been argued to break down for tissue, where a single crop typically contains several heterogeneous structures, and to cause models to encode stain texture rather than biological organization [14].

Modeling long-range dependencies between tiles is then left to downstream aggregation, commonly multiple instance learning (MIL). Attention-based, hierarchical, and graph-based variants do improve task-level performance [14], yet they operate on frozen representations from an encoder that was not optimized to capture these dependencies. Tizhoosh [14] argues that such aggregation can compensate for deficiencies in the upstream representation while masking them. Since any aggregator is bounded by the information present in these representations, we argue that long-range dependencies should instead be learned during initial pretraining.

The value of moving beyond isolated tile representations is also evident from approaches that incorporate spatially structured context during or after pretraining. For example, Quinones Robles et al. [18] add positive pairs formed from spatially adjacent tiles to several invariance-based self-supervised learning (SSL) methods, including Barlow Twins [19], BYOL [20], VICReg [21] and DINOv2 [16], in addition to the usual augmentation-based pairs, improving both tile- and slide-level classification in most settings. Since tiles farther apart are increasingly likely to belong to different tissue classes, the authors restrict contextual pairs to immediate neighbors, and downstream gains are correspondingly highest for encoders pretrained with adjacent tiles and fall off with distance. This confines the spatial range over which such pairs can usefully incorporate context, since they must remain visually and semantically similar to serve as positives. TICON [22] instead keeps the tile encoder frozen and pretrains a transformer to reconstruct masked tile representations from the remaining, unmasked ones inside a larger context of 16×16 tiles, extrapolating to the full slide at inference. This produces a context-enriched representation for every tile and likewise improves tile- and slide-level classification. Together, these approaches expose complementary limitations: the former incorporates spatial context during pretraining, but only through direct pairwise relations between sufficiently similar tiles, whereas the latter introduces broader context only after tile-level encoder pretraining. Neither addresses the more general limitation of single-step objectives, which do not model long-range dependencies through a sequence of intermediate crops during initial pretraining from pixels.

An alternative is to model these dependencies sequentially during pretraining, while explicitly conditioning on the transformations that connect successive

crops. A recent self-supervised method, seq-JEPA [23], processes a sequence of crops, each paired with the action relating it to the next, and aggregates them into a single representation. From this aggregate representation and the upcoming action, it then predicts the representation of the final crop in the sequence. This removes the similarity requirement: two regions need not be similar, only predictable from one another given the action between them. Relative transformations between crops are also better preserved than by invariance-trained encoders [23]. Because the context consists of a chain of intermediate crops and actions rather than a single pair, predictions can depend on information accumulated across the sequence rather than only on a direct relation between two crops. At the same time, the transformations relating successive crops can compose across the sequence. This is reflected in path integration, where the cumulative transformation from the first to the last crop can be decoded from the aggregate representation together with the final action embedding [23]. Beyond composition, a sequence can accumulate context from a limited observation budget rather than from all observations at once, a property relevant to whole-slide images, whose exhaustive encoding at high resolution can become computationally expensive. This approach is more closely aligned with biological vision, which acquires context through a sequence of saccades, each accompanied by an internal copy of the motor command leading to the next observation, rather than from a single observation [24].

We therefore argue that sequential, action-conditioned prediction points toward an alternative direction for biomedical self-supervised models. As a proof of concept, we adapt seq-JEPA [23] to 256×256 px histopathology images and evaluate both the original formulation and our per-timestep variant against a SimCLR baseline on tissue classification and held-out-crop retrieval.

Our contributions are as follows.

- **We introduce a variable-scale extension of seq-JEPA’s original saccade setting**, adding a zoom dimension to its 2D shift actions so that successive crops can differ in scale as well as position. While seq-JEPA’s separate augmentation setting already encodes crop position, height, and width independently as actions, our formulation retains the saccade setting’s 2D shift structure while leaving photometric augmentations

outside the action space. This yields a hybrid setting in which seq-JEPA and SimCLR share the same per-crop augmentation pipeline.

- **We propose seq-JEPA-dense**, extending seq-JEPA’s action-conditioned prediction objective from a single prediction of the final representation to next-representation prediction at every timestep under causal masking. This yields dense next-representation supervision, as also explored in recent sequential image SSL methods [25–27], while retaining seq-JEPA’s explicit relative-action conditioning: each token carries its incoming action, and the outgoing action is supplied to the predictor at every step.

2 Related Work

Sequential prediction has a long history as a self-supervised learning objective for static images. Contrastive Predictive Coding [28] recasts it as a general-purpose pretraining objective, summarizing the patches seen so far with an autoregressive model and predicting the representations of future patches contrastively. Within this line, AIM [29] and NEPA [25] retain an autoregressive structure but replace the contrastive objective with direct regression, predicting future patches in pixel and latent space respectively. NEPA predicts the next patch representation at each sequence position under causal masking, using a cosine-similarity loss averaged across positions. These approaches, however, traverse a fixed grid of patches in the same order for every image, typically row by row from top to bottom, so that the next target is given by the order rather than specified by a transformation.

In contrast, I-JEPA [7] predicts the representations of held-out target regions directly from the visible image context, with each target predicted independently rather than as part of a sequence. DSeq-JEPA [26] introduces sequential prediction into this framework, using a saliency map derived from the target encoder’s attention to both select the target regions and define a sequential prediction order. The target-region representations are then predicted sequentially along this content-dependent order. While this yields a content-dependent rather than fixed prediction order, each target is still specified by an absolute positional

token rather than by a relative transformation from the preceding region. DSeq-JEPA outperforms an I-JEPA baseline across several vision benchmarks, with ablations indicating that the improvement depends on combining sequential prediction with discriminative region selection [26].

Image World Models (IWM) [30] are a complementary extension of I-JEPA. IWM retain I-JEPA-style mask tokens, whose positional embeddings specify where target representations are to be predicted, while conditioning the predictor on known photometric and corruption transformations between source and target views. Depending on the conditioning and predictor capacity, the learned representation can then retain transformation-dependent information rather than becoming invariant to those transformations. The spatial component of the prediction task, however, is still specified by absolute target positions rather than by relative transformations between views, and prediction remains pairwise rather than sequential.

seq-JEPA [23], which forms the basis of our work, brings together transformation conditioning and sequential prediction. In its saccade setting, the observation sequence is derived from DeepGaze IIE [31], a neural model for predicting human visual saliency and fixation distributions, rather than from a fixed raster traversal or the target encoder’s attention. Fixations are drawn from this saliency map with inhibition of return, which suppresses already visited regions so that successive crops overlap less and the sequence is less redundant. Each observation is paired with the relative transformation, such as a change in position, that connects it to the succeeding one. A transformer aggregator summarizes this history of observation–action pairs, and a predictor then predicts the representation of the next observation from the aggregate conditioned on the upcoming action. This formulation allows the encoder to retain observation–action-equivariant information while the aggregated representation remains suitable for invariant downstream tasks. Following [23], equivariance means that for each action a relating two views x_1 and x_2 of an image, and encoder f , there exists a transformation u_a in latent space such that $f(x_2) = u_a(f(x_1))$. When u_a is the identity, this reduces to invariance, $f(x_2) = f(x_1)$.

Related self-supervised methods in biomedical imaging have explored the same ingredients in different combinations. CheXWorld [32] predicts target-region

representations from their relative displacement in chest X-ray images among other self-supervised objectives, while X-WIN [33] predicts representations of novel X-ray projections conditioned on changes in viewpoint around a corresponding computed tomography (CT) volume. Both explicitly model transformations between observations, but formulate prediction as a single-step relation rather than over an extended sequence.

GazeWorld [27] instead introduces sequential prediction in a biomedical setting. It refines a pretrained encoder by autoregressively predicting representations along expert gaze trajectories. In its next-fixation branch, however, the location of the next fixation is not supplied to the predictor and must be inferred from the fixation history, while an equally weighted spatial-completion branch specifies unvisited target patches by their absolute positions, as in I-JEPA. It therefore models an extended observation sequence while relying on pretrained encoder representations rather than learning them from scratch.

Wang and Chen [34] combine sequential and action-conditioned prediction for volumetric magnetic resonance imaging (MRI) data. They render a fixed 3D volume as a sequence of 2D slices accompanied by control vectors describing parameters such as position, orientation, and scale, and train a causal dynamics model to predict representations along the resulting trajectory. To our knowledge, this is the closest prior formulation in a biomedical context to our setting. It nevertheless differs in two important respects. First, the visual representations are learned in a separate masked-reconstruction stage and frozen before dynamics training, so the sequential objective shapes only the dynamics model and not the encoder. Second, its actions encode absolute per-frame rendering states rather than relative transformations between successive observations. Relative transformations, in contrast, do not require a global reference frame, as each action is defined only with respect to the preceding observation.

As discussed above, prior work has explored both sequential prediction in latent space and explicit transformation conditioning as self-supervised objectives for static images, separately and in combination. On natural images, IWM condition prediction on the transformation between pairs of views, while seq-JEPA combines explicit inter-view actions with sequential prediction. In biomedical

imaging, transformation-conditioned methods remain predominantly single-step, while sequential approaches either leave the upcoming target implicit or specify it by absolute positions or acquisition states, or build on pretrained or separately learned representations. Our work studies sequential self-supervised representation learning in histopathology conditioned on the relative changes in position and scale between successive crops.

3 Methods

3.1 Dataset

We use the PathoROB Tolkach ESCA dataset [35, 36], comprising 16,300 histopathology images of size 256×256 px across four cohorts and six tissue classes (Figure 1), at an approximate physical resolution of $0.75 \mu\text{m}/\text{px}$ [35]. We pool all four cohorts and construct a train/test split for the present study, which differs from the original PathoROB evaluation protocol. We use a five-fold stratified group split, stratifying by tissue class and grouping by `slide_id`, with shuffling and random seed 42; the first fold is used as the test set. This yields 13,400 training images (82.2%) and 2,900 test images (17.8%). Of the 77 slide/case identifiers, 62 are assigned to training and 15 to testing, with no `slide_id` shared between the two sets.

3.2 Augmentation Strategy

In seq-JEPA’s original saccade setting, actions encode only the relative displacement $\delta_i = (\Delta u_i, \Delta v_i) \in \mathbb{R}^2$ between fixed-size crops, where $(u_i, v_i) \in [0, 1]^2$ denotes the center of crop i (row, column) normalized by the source-image size, whereas its augmentation setting additionally represents crop height and width separately in the action space. We extend the saccade setting to variable-scale crops by adding the change in zoom as an additional action dimension, where the zoom of crop i with side length l_i in source-image pixels is $\zeta_i = \log(\text{input_size}/l_i)$, with `input_size` = 128 px the fixed encoder input size to which every sampled crop is resized. This results in $\delta_i = (\Delta u_i, \Delta v_i, \Delta \zeta_i) =$

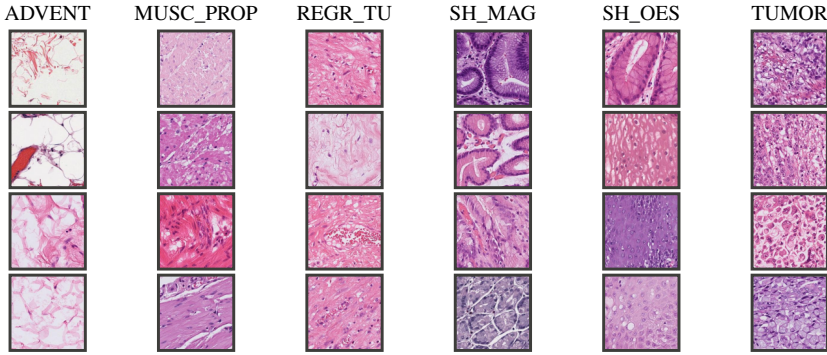


Figure 1: Example images of the six tissue classes in the PathoROB Tolkach ESCA dataset [35, 36]. Each column corresponds to one tissue class and contains four randomly sampled images. Class abbreviations follow [35]: ADVENT = adventitial tissue; MUSC_PROP = muscularis propria; REGR_TU = regression areas; SH_MAG = gastric mucosa; SH_OES = oesophageal mucosa; TUMOR = vital tumor tissue.

$(u_{i+1} - u_i, v_{i+1} - v_i, \zeta_{i+1} - \zeta_i) \in \mathbb{R}^3$, where $\Delta\zeta_i = \log(l_i/l_{i+1})$ is independent of input_size and positive when zooming in. The logarithmic parameterization makes multiplicative changes in zoom compose additively across successive crops. Thus, the actions for seq-JEPA encode changes in crop location and scale, but not the photometric augmentations applied to the crop. This design allows for SimCLR and seq-JEPA to share an identical per-crop augmentation pipeline. From each 256×256 px image, square crops covering 8%–100% of the source-image area are sampled, resized by bicubic interpolation to 128×128 px (input_size), and individually augmented with color jitter ($p = 0.8$), random grayscale ($p = 0.2$), and Gaussian blur ($p = 0.5$; 25×25 kernel, $\sigma \in [0.1, 2.0]$). SimCLR samples $S = 2$ such crops and uses them as a positive contrastive pair. seq-JEPA samples $S = 5$ crops, of which the first $N = 4$ are observed together with the action embeddings of their $(\Delta u_i, \Delta v_i, \Delta \zeta_i)$ transformations, while the last is held out as the prediction target. Crops are sampled at random and independently of the model’s state. Unlike the original saccade setting [23], crop centers are not guided by a saliency map or inhibition of return, crops always lie fully inside the image, and each crop additionally receives photometric augmentation. For the test split and for visualization, photometric augmentation is disabled and crop sampling is seeded per image, so that all models are

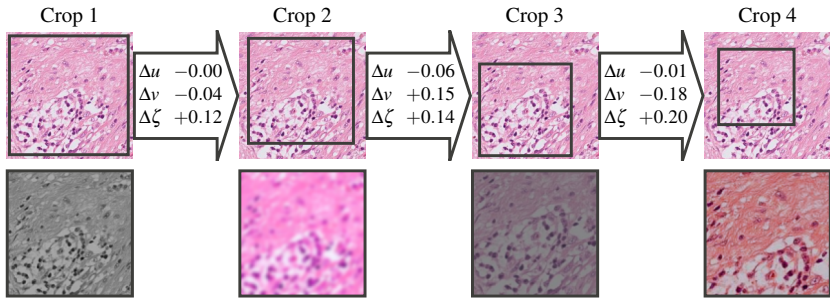


Figure 2: Example sequence of four crops sampled from a single pathology image. **Top:** the source image with the bounding box for each crop outlined. **Bottom:** the corresponding extracted and augmented crop. Arrows show the transform between consecutive crops, given as the change in position and log-zoom, $(\Delta u, \Delta v, \Delta \zeta)$.

evaluated on identical crops. An example sequence of augmented crops is shown in Figure 2.

3.3 Models

We compare three models: A SimCLR [1] baseline, seq-JEPA [23], and seq-JEPA-dense, which replaces the single loss calculation at the end of the sequence with supervision for every consecutive step. All three methods use a ResNet-18 [37] backbone with random initialization, yielding crop representations $z_i \in \mathbb{R}^{512}$.

SimCLR. The encoder is followed by a 3-layer multilayer perceptron (MLP) projection head (2048-2048-2048 hidden units, batch-normalized), trained with the NT-Xent contrastive loss at temperature $\tau = 0.5$ on the two augmented crops described in Section 3.2. This follows the SimCLR baseline from the original seq-JEPA implementation rather than SimCLR’s ImageNet defaults ($\tau = 0.1$, two layers, 128-d).

seq-JEPA. Following the original setup, a sequence of $N = S - 1$ encoded crops, each concatenated with an embedding $a_i \in \mathbb{R}^{128}$ of its *outgoing* action, with a zero vector in place of the action for the last crop, is aggregated by a lightweight transformer encoder (4 heads, 3 layers) via a learnable [AGG] token $\in \mathbb{R}^{640}$.

The aggregator uses no positional encoding. The aggregate representation at the [AGG] token, $z_{\text{AGG}} \in \mathbb{R}^{640}$, is concatenated with the final action embedding and passed through a two-layer MLP predictor (hidden dimension 1024, output dimension 512) to predict the encoder’s representation of the held-out final crop, with a stop-gradient applied to the target and trained with a negative cosine similarity loss.

seq-JEPA-dense. This variant shares the same backbone, aggregator, and predictor architecture as seq-JEPA but predicts at every position in the sequence instead of only the last, using a causal attention mask so that each token only attends to itself and preceding tokens. The prepended [AGG] token is retained from seq-JEPA but is no longer used as an aggregate readout. Instead, it serves as a learned beginning-of-sequence token [BOS] that provides a common initial context for the causally masked sequence. Each token now carries its *incoming* relative action, with a zero vector for the first crop, which has no incoming action, so that the action pointing at a given position’s target is withheld from the aggregator and injected only at the predictor. The predictor is applied at each position, conditioning that position’s causally-attended transformer output $h_i \in \mathbb{R}^{640}$ on its own outgoing action to forecast the next crop’s representation, using teacher forcing, i.e., conditioning on the encoded observed crops rather than on earlier predictions. Targets are the encoder’s representations at every step with a stop-gradient applied, and the loss averages the negative cosine similarity over all N predictions, yielding denser supervision per training step.

The prediction objectives of seq-JEPA and seq-JEPA-dense are illustrated in Figure 3. The official seq-JEPA code [23] supports both exponential moving average (EMA) and online targets. Unlike the original paper, which uses an EMA target encoder (decay 0.996), we use the online-target option, computing targets with the online encoder under stop-gradient. SimSiam [38] showed that stop-gradient with a predictor can prevent collapse without a momentum encoder. This choice also keeps all compared methods free of a momentum encoder and, for seq-JEPA-dense, lets targets reuse the observed crop encodings, saving $N - 1$ encoder passes per sequence.

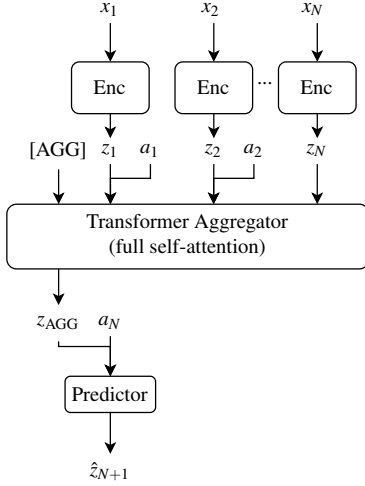
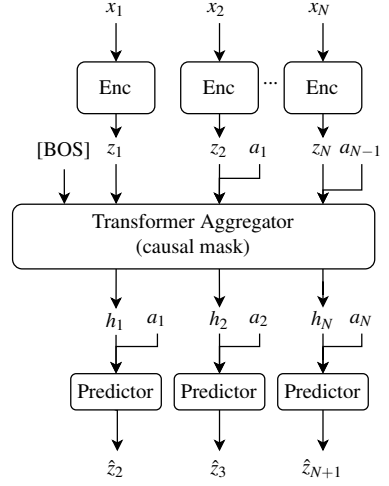
(a) seq-JEPA*Final-step prediction***(b) seq-JEPA-dense***Per-timestep prediction*

Figure 3: Prediction scheme of (a) seq-JEPA [23] and (b) seq-JEPA-dense. a_i denotes the relative action embedding from crop x_i to x_{i+1} . In (a), the aggregator attends over (z_i, a_i) for $i = 1, \dots, N-1$ and $(z_N, \mathbf{0})$, and the predictor uses the [AGG] output z_{AGG} , conditioned on a_N , to predict z_{N+1} . In (b), a causal mask is applied, and each token instead carries its incoming action (z_i, a_{i-1}) , with $a_0 = \mathbf{0}$; the predictor uses the output h_i , computed from [BOS] and the tokens $(z_1, \mathbf{0}), (z_2, a_1), \dots, (z_i, a_{i-1})$, conditioned on a_i , to predict the encoder representation of the next crop z_{i+1} . The aggregate token [AGG] is repurposed as a beginning-of-sequence token [BOS], and the representation used for the classification task is read from h_N instead of z_{AGG} . All Enc blocks denote the same encoder, and all Predictor blocks the same predictor.

3.4 Experimental Setup

3.4.1 Training

SimCLR, seq-JEPA, and seq-JEPA-dense are pretrained from scratch on the training split with the same augmentation pipeline and epoch budget. We use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of 4×10^{-4} , weight decay of 10^{-3} , and batch size of 512. The learning rate is linearly warmed up from 10^{-5} over 20 epochs and kept constant thereafter. All models are trained for 1,000 epochs in bf16 mixed precision with a fixed seed on a single NVIDIA RTX 3090 GPU, and the final checkpoint is evaluated.

3.4.2 Classification Evaluation

After pretraining, we freeze each model and train linear classification probes for 100 epochs using Adam with a learning rate of 10^{-3} . The probes are trained on the training split and evaluated by top-1 accuracy (percentage of test samples correctly classified) on the 6-class classification task corresponding to the six tissue classes.

For each method, we train two linear classification probes. The *res* probe operates on the single-crop ResNet representation, while the *agg* probe operates on a multi-crop representation aggregated over $N = 4$ crops. For SimCLR, which has no learned aggregator, we mean-pool the encoder representations across the four crops as a parameter-free aggregation. For seq-JEPA, we use the z_{AGG} token; for seq-JEPA-dense, we use the transformer output at the last crop token, h_N .

3.4.3 Retrieval Evaluation

We evaluate retrieval with Hit@1, the fraction of queries for which the correct target is the top-ranked candidate. This requires recovery of the specific held-out crop rather than any crop with the same tissue label. For each test image, we sample $S = 5$ crops as during training: $N = 4$ context crops and one held-out target. Candidates are ranked by the cosine similarity between the query and their encoder representations. For both seq-JEPA variants, the query is the predictor output conditioned on the action leading to the target, taking the final-step prediction in the dense case. For SimCLR, the query is the mean-pooled encoder representations of the four context crops, matching the classification evaluation *agg* probe.

Our primary setting, within-image retrieval, adapts the same-instance retrieval protocol of Garrido et al. [39], also used by Ghaemi et al. [23] on the 3DIEBench dataset [39]. We retain the same-instance restriction, but tailor the candidate set to our sequential setting, using the four observed context crops as distractors alongside the held-out target ($N + 1 = 5$; chance 20%). Because all candidates

come from the same image, slide-, stain-, and center-specific cues are held constant, so success requires identifying the requested crop. SimCLR is excluded because its query averages the context crops and is not action-conditioned.

The secondary setting enlarges the candidate set: we keep one held-out target crop per test image, resulting in $M = 2,900$ crops, and rank each query against all of them (chance $\approx 0.034\%$). A hit thus requires identifying the correct *image*, not the correct crop within a known image. Since the distractors now come from other images, they can differ from the target in stain and slide-level appearance, not only in morphology. Hit@1 therefore cannot separate morphological specificity from those cues in this setting.

3.4.4 Representation Space Analysis

To qualitatively examine the learned representation spaces, we visualize the *res* and *agg* representations of each model using Uniform Manifold Approximation and Projection (UMAP) [40] over all 16,300 images in the dataset. The *res* representation corresponds to the encoder output of a single crop ($N = 1$), while *agg* corresponds to the aggregated representation over $N = 4$ crops, as used for the respective classification probes. We fit UMAP separately to each model and representation using Euclidean distance on the unnormalized feature vectors, with 30 nearest neighbors, a minimum distance of 0.1, and a fixed random seed of 42. Points in the resulting two-dimensional projections are colored by tissue class.

4 Results

4.1 Classification Results

Table 1 shows that SimCLR achieves the highest top-1 accuracy in both settings, followed by seq-JEPA-dense and seq-JEPA. Relative to seq-JEPA, seq-JEPA-dense improves accuracy by 3.3 and 3.6 percentage points in the single-crop and aggregated settings, narrowing the gap to SimCLR from 5.1 to 1.8 and from 5.0 to 1.4 percentage points.

Table 1: Linear-probe accuracy for tissue classification across six classes on the test split used in this study for PathoROB Tolkach ESCA (2,900 images). Results are shown for SimCLR, seq-JEPA, and seq-JEPA-dense, comparing single-crop and aggregated evaluation across $N = 4$ crops, reported as top-1 accuracy.

Method	Crops (N)	Top-1 Acc. (%)
SimCLR (<i>res</i>)	1	91.7
seq-JEPA (<i>res</i>)	1	86.6
seq-JEPA-dense (<i>res</i>)	1	89.9
SimCLR (<i>agg</i>)	4	93.3
seq-JEPA (<i>agg</i>)	4	88.3
seq-JEPA-dense (<i>agg</i>)	4	91.9

4.2 Retrieval Results

In the within-image retrieval setting (Table 2), both seq-JEPA variants exceed chance (42.1% and 41.8% against 20%), with comparable Hit@1. In the cross-image retrieval setting, seq-JEPA-dense identifies the held-out crop most often (95.3%), ahead of SimCLR (87.0%) and seq-JEPA (80.6%).

4.3 Representation Space Analysis

Figure 4 shows UMAP projections of the *res* ($N = 1$) and *agg* ($N = 4$) representations for SimCLR, seq-JEPA, seq-JEPA-dense, colored by tissue class. SimCLR’s clusters appear most visually separated of the three methods, forming distinct, well-isolated groups in both the *res* and *agg* settings. Among seq-JEPA and seq-JEPA-dense, the relative separation depends on the setting: at $N = 1$, seq-JEPA-dense’s clusters appear more visually separated than seq-JEPA’s, while at $N = 4$ this pattern reverses, with seq-JEPA’s clusters appearing more separated.

Table 2: Held-out-crop retrieval on the test split used in this study for PathoROB Tolkach ESCA (2,900 images), reported as Hit@1. Within-image: the target is ranked against the four context crops from the same image (5 candidates). Cross-image: the target is ranked against the held-out targets of all $M = 2,900$ test images. SimCLR is not action-conditioned and is therefore evaluated only in the cross-image setting.

Method	Within-image (%)	Cross-image (%)
<i>Chance</i>	<i>20.0</i>	<i>0.034</i>
SimCLR	–	87.0
seq-JEPA	42.1	80.6
seq-JEPA-dense	41.8	95.3

5 Discussion

In our experiments, both seq-JEPA variants support action-conditioned retrieval within an image, and seq-JEPA-dense produces the most image-specific retrieval queries among the compared methods while approaching the contrastive baseline in tissue classification.

seq-JEPA-dense narrows the classification gap to SimCLR relative to seq-JEPA (Table 1). As it also differs from seq-JEPA in its attention mask, the assignment of actions to tokens, and its readout, this gain cannot be attributed to denser supervision alone. The UMAP projections, however, show less separated aggregate clusters for seq-JEPA-dense than for seq-JEPA, the reverse of their classification ordering; apparent separation in these 2D projections is thus not a reliable proxy for linear separability in the feature space. SimCLR nevertheless remains strongest for tissue classification. A plausible explanation is our single-image setting: with all crops drawn from one 256×256 px image and an input resolution of 128×128 px, a single crop can already capture most of the image’s information, so the sequence adds little class-relevant information beyond its largest crop. An advantage of sequential, action-conditioned prediction over single-step objectives such as SimCLR would therefore be expected mainly when the relevant content cannot be captured by any single crop, as in large whole-slide images.

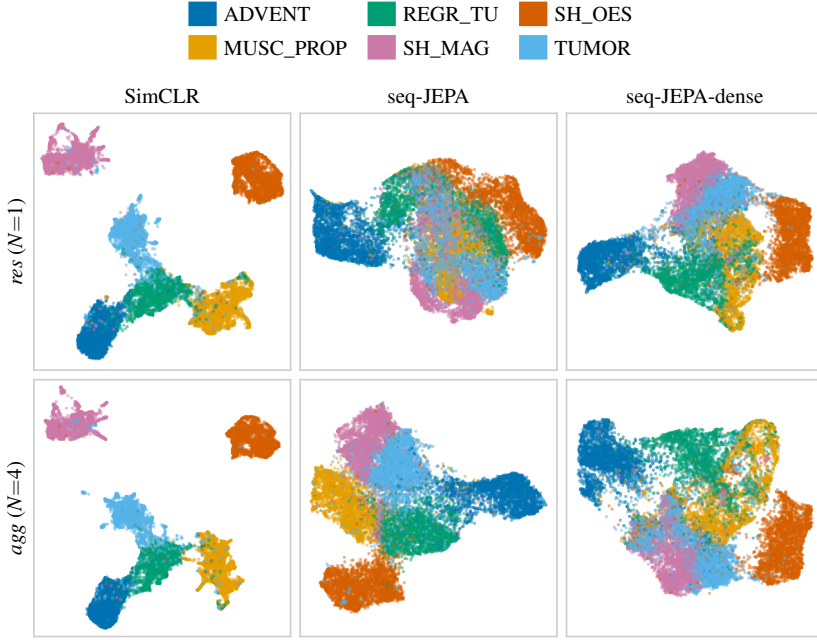


Figure 4: UMAP projections of the learned representations of SimCLR, seq-JEPA, and seq-JEPA-dense for all 16,300 images in the PathoROB Tolkach ESCA dataset [35, 36], colored by tissue class. Top row: *res* representations (encoder output for a single crop, $N = 1$). Bottom row: *agg* representations (aggregated over $N = 4$ crops).

Recovering a crop specified by an action requires a query that encodes the requested relative action. Both seq-JEPA variants exceed chance substantially within an image (42.1% and 41.8% against 20% chance rate), indicating that the predictor uses the queried action rather than returning an image-level average, although this evaluation does not show whether it relies on position, zoom, or both. seq-JEPA-dense provides no advantage in this setting over seq-JEPA despite its classification gain. We hypothesize that this results from the redundancy of the sampled sequence: crops are drawn independently and may cover up to 100% of the image, whereas seq-JEPA’s saccade setting uses inhibition of return to avoid overlapping, redundant sequences. Consequently, a representation similar to that of the held-out crop may be constructed largely from the preceding crop representations, reducing pressure on the predictor

to rely on the relative actions. An inhibition-of-return-like constraint on crop overlap is difficult to impose here, as on 256×256 px images it would require restricting crops to small areas and thus depart from the crop distribution shared with SimCLR.

In the cross-image retrieval task, seq-JEPA-dense identifies the held-out crop more often than SimCLR, the reverse of the classification ordering. Because candidate images span medical centers, success requires only that the query is specific to the source image, and slide- or center-specific batch effects may contribute. seq-JEPA-dense matches seq-JEPA within an image but reaches a considerably higher cross-image Hit@1 (95.3% vs. 80.6%), indicating that it retains more image-specific information. This may stem from the denser training signal, but the accompanying architectural changes were not isolated. The matching within-image results do not show that this additional information is non-positional: because redundant crop sequences could allow the target to be approximated largely without the action, within-image retrieval may be only weakly sensitive to positional information. Both SimCLR and the seq-JEPA variants encourage image-specific representations: SimCLR explicitly discriminates between crops of different images, whereas the seq-JEPA variants predict a crop from other crops of the same image, which requires image-specific information. One interpretation of seq-JEPA-dense ranking above SimCLR is that its query matches the target crop more specifically: the SimCLR query averages the context crops and thus represents the image as a whole, whereas seq-JEPA-dense predicts the particular target crop from the requested action, which may help to distinguish it from similar images of the same slide. How much of the preserved signal is morphological, and therefore potentially useful for tasks that depend on local structure, or technical, cannot be determined from this evaluation alone.

6 Limitations

Our evaluation is limited to a single dataset comprising 13,400 training images. Although the dataset splits and training and evaluation runs are seeded for

reproducibility, we train only a single run per method and do not perform cross-validation. Consequently, we cannot report variability across random seeds for either classification or retrieval performance.

Moreover, context and action do not fully determine the content of non-overlapping crops, yet the deterministic predictor outputs a single point estimate, likely approximating an average over plausible targets rather than any specific one.

While we aim to match the training settings of SimCLR and the seq-JEPA variants as closely as possible, the methods differ in the amount and type of self-supervision received per loss update. SimCLR operates on two augmented crops per update, whereas the seq-JEPA variants use five. Moreover, seq-JEPA-dense receives a denser training signal by predicting each consecutive step. Conversely, SimCLR benefits from explicit in-batch negatives as part of its contrastive objective, while the seq-JEPA variants carry a larger parameter count from the transformer aggregator their objective requires. These differences make a strictly controlled comparison difficult and should be considered when interpreting relative performance across the methods. The comparison is also limited to a single baseline family (contrastive, SimCLR); non-contrastive self-distillation (e.g., BYOL [20], DINO [3]) and reconstruction-based objectives (e.g., MAE [5]) are outside the scope of this proof-of-concept study.

7 Conclusion and Future Work

We adapt seq-JEPA to computational pathology by extending its action space with a zoom dimension, and introduce seq-JEPA-dense, a variant that predicts at every position in the crop sequence under a causal mask. In a single pretraining run per method, seq-JEPA-dense improves linear-probe accuracy over seq-JEPA by 3.3 and 3.6 percentage points in the single-crop and aggregated settings, narrowing the gap to a SimCLR baseline to 1.8 and 1.4 points. Both variants recover a held-out crop within its own image at roughly twice the chance rate, in a setting where slide-, stain-, and center-specific cues are constant by construction. In the cross-image retrieval setting, seq-JEPA-dense additionally retrieves the held-out crop more accurately than both seq-JEPA and SimCLR,

suggesting that seq-JEPA-dense preserves image-specific information not fully reflected by linear-probe classification.

Beyond the single-image setting evaluated here as a proof of concept, sequential, action-conditioned prediction could extend to biomedical images whose informative content spans large areas or volumes, such as whole-slide images, microscopy z-stacks, or volumetric scans. For whole-slide images, a trajectory could move across both position and magnification through the same shift and zoom actions used here, allowing the model to relate fine-grained local morphology to tissue organization over much larger spatial scales. In this way, the same predictive objective could capture long-range dependencies and local details jointly, rather than treating tile-level representation learning and slide-level aggregation as separate stages. Scaling the approach to gigapixel slides would, however, require more efficient handling of long crop sequences and more efficient crop encoders. Furthermore, the crop trajectory would need to be sampled such that subsequent crops are sufficiently predictable for the model to learn meaningful representations. A simple sampling policy could be to sample crops that are close in position and zoom level, i.e., with small relative action magnitude. Alternatively, the sampling policy could itself be learned without supervision, e.g., by selecting actions that minimize the uncertainty of the aggregated representation. On small-scale active vision tasks, Sharafeldin et al. [41] show that such uncertainty-driven sampling learns to focus on informative image regions and improves downstream classification accuracy compared to random sampling. Adapting such an approach to seq-JEPA and to whole-slide images is left for future work.

A second direction is task-specific adaptation of the sampling policy after initial pretraining. On whole-slide images, random sampling is unlikely to provide an efficient exploration strategy for downstream tasks, making it important to learn which regions and magnifications should be queried for a given task. One possibility is supervised adaptation, for example by freezing the pretrained encoder and training an action-selection policy from downstream labels or region-level supervision so that the resulting crop sequence becomes task-directed. The causal, action-conditioned formulation also lends itself naturally to reinforcement learning, where the policy could instead be optimized directly for downstream performance.

Segmentation could be an appropriate next downstream evaluation task. The retrieval results suggest that seq-JEPA-dense retains information that classification probes do not reward, and segmentation requires spatially resolved predictions, whereas the tissue-classification probe produces one label per image. Action-conditioned world models have been applied to biomedical image segmentation before [32, 34] (Section 2), though not with relative actions over an observation sequence.

8 Code Availability

The code for all experiments is available at <https://github.com/lischilpp/seq-jepa-dense>.

9 Acknowledgments

The present contribution is supported by the Helmholtz Association under the program “Natural, Artificial and Cognitive Information Processing (NACIP)” and the joint research school “Helmholtz Information & Data Science School for Health (HIDSS4Health)”.

During the preparation of this work, the authors used the AI language models Claude Sonnet 5, Claude Opus 5 and 5.5 (Anthropic) and GPT-5.6 Sol (OpenAI) to assist with literature search, software development, and drafting and revising the manuscript text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of this manuscript.

References

- [1] T. Chen *et al.*, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research,

- H. Daumé III and A. Singh, Eds., vol. 119. PMLR, Jul. 2020, pp. 1597–1607.
- [2] S. Azizi *et al.*, “Big self-supervised models advance medical image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3478–3488.
 - [3] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9650–9660.
 - [4] F. Pérez-García *et al.*, “Exploring scalable medical image encoders beyond text supervision,” *Nature Machine Intelligence*, vol. 7, no. 1, pp. 119–130, Jan. 2025.
 - [5] K. He *et al.*, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 16 000–16 009.
 - [6] O. Kraus *et al.*, “Masked autoencoders for microscopy are scalable learners of cellular biology,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 11 757–11 768.
 - [7] M. Assran *et al.*, “Self-supervised learning from images with a joint-embedding predictive architecture,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 15 619–15 629.
 - [8] A. A. H. Khan *et al.*, “RadJEPa: Radiology encoder for chest X-rays via joint embedding predictive architecture,” arXiv:2601.15891, 2026.
 - [9] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, Jul. 2021, pp. 8748–8763.

- [10] S. Zhang *et al.*, “A multimodal biomedical foundation model trained from fifteen million image–text pairs,” *NEJM AI*, vol. 2, no. 1, p. AIoa2400640, 2025.
- [11] M. Gadermayr and M. Tschuchnig, “Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations and future potential,” *Computerized Medical Imaging and Graphics*, vol. 112, p. 102337, 2024.
- [12] V. van Veldhuizen *et al.*, “Foundation models in medical imaging: A review and outlook,” arXiv:2506.09095, 2025.
- [13] A. H. Song *et al.*, “Artificial intelligence for digital and computational pathology,” *Nature Reviews Bioengineering*, vol. 1, no. 12, pp. 930–949, Oct. 2023.
- [14] H. R. Tizhoosh, “Rethinking foundation models in pathology,” *Nature Biomedical Engineering*, vol. 10, no. 8, pp. 1513–1518, Jun. 2026.
- [15] M. Caron *et al.*, “Unsupervised learning of visual features by contrasting cluster assignments,” in *Advances in Neural Information Processing Systems*, H. Larochelle *et al.*, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9912–9924.
- [16] M. Oquab *et al.*, “DINOv2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research*, 2024.
- [17] O. Siméoni *et al.*, “DINOv3,” arXiv:2508.10104, 2025.
- [18] W. R. Quinones Robles *et al.*, “Leveraging spatial context for positive pair sampling in histopathology image representation learning,” arXiv:2503.05170, 2026.
- [19] J. Zbontar *et al.*, “Barlow Twins: Self-supervised learning via redundancy reduction,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, Jul. 2021, pp. 12 310–12 320.

- [20] J.-B. Grill *et al.*, “Bootstrap Your Own Latent - a new approach to self-supervised learning,” in *Advances in Neural Information Processing Systems*, H. Larochelle *et al.*, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 271–21 284.
- [21] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-invariance-covariance regularization for self-supervised learning,” in *International Conference on Learning Representations*, 2022.
- [22] V. Belagali *et al.*, “TICON: A slide-level tile contextualizer for histopathology representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, Jun. 2026, pp. 5440–5451.
- [23] H. Ghaemi, E. B. Muller, and S. Bakhtiari, “seq-JEPA: Autoregressive predictive learning of invariant-equivariant world models,” in *Advances in Neural Information Processing Systems*, D. Belgrave *et al.*, Eds., vol. 38. Curran Associates, Inc., 2025, pp. 32 943–32 973.
- [24] R. H. Wurtz, “Corollary discharge contributions to perceptual continuity across saccades,” *Annual Review of Vision Science*, vol. 4, pp. 215–237, 2018.
- [25] S. Xu *et al.*, “Next-embedding prediction makes strong vision learners,” arXiv:2512.16922, 2025.
- [26] X. He *et al.*, “DSeq-JEPA: Discriminative sequential joint-embedding predictive architecture,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.
- [27] Y. Li *et al.*, “A world model of radiologist reading for medical image representation learning,” arXiv:2605.23992, 2026.
- [28] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv:1807.03748, 2019.
- [29] A. El-Nouby *et al.*, “Scalable pre-training of large autoregressive image models,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research,

- R. Salakhutdinov *et al.*, Eds., vol. 235. PMLR, Jul. 2024, pp. 12 371–12 384.
- [30] Q. Garrido *et al.*, “Learning and leveraging world models in visual representation learning,” arXiv:2403.00504, 2024.
 - [31] A. Linardos *et al.*, “DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 12 919–12 928.
 - [32] Y. Yue *et al.*, “CheXWorld: Exploring image world modeling for radiograph representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 20 778–20 788.
 - [33] Z. Yang *et al.*, “X-WIN: Building chest radiograph world model via predictive sensing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2026, pp. 6920–6930.
 - [34] Y. Wang and Q. Chen, “3D MRI image pretraining via controllable 2D slice navigation task,” arXiv:2605.06487, 2026.
 - [35] J. Kömen *et al.*, “Towards robust foundation models for digital pathology,” *Nature Communications*, vol. 17, no. 1, Jun. 2026.
 - [36] Y. Tolkach *et al.*, “Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study,” *The Lancet Digital Health*, vol. 5, no. 5, pp. e265–e275, 2023.
 - [37] K. He *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.
 - [38] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15 750–15 758.

- [39] Q. Garrido, L. Najman, and Y. LeCun, “Self-supervised learning of split invariant equivariant representations,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause *et al.*, Eds., vol. 202. PMLR, Jul. 2023, pp. 10 975–10 996.
- [40] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” arXiv:1802.03426, 2020.
- [41] A. Sharafeldin, N. Imam, and H. Choi, “Active sensing with predictive coding and uncertainty minimization,” *Patterns*, vol. 5, no. 6, p. 100983, Jun. 2024.