Learning Sparse Visual Representations via Spatial-Semantic Factorization

Theodore Zhengde Zhao [1] **Sid Kiblawi** [1] **Jianwei Yang** [2 3] **Naoto Usuyama** [1] **Reuben Tan** [1] **Noel C Codella** [1] **Tristan Naumann** [1] **Hoifung Poon** [1] **Mu Wei** [1]

Abstract

Self-supervised learning (SSL) faces a fundamental conflict between semantic understanding and image reconstruction. High-level semantic SSL (e.g., DINO) relies on global tokens that are forced to be location-invariant for augmentation alignment, a process that inherently discards the spatial coordinates required for reconstruction. Conversely, generative SSL (e.g., MAE) preserves dense feature grids for reconstruction but fails to produce high-level abstractions. We introduce STELLAR, a framework that resolves this tension by factorizing visual features into a low-rank product of semantic concepts and their spatial distributions. This disentanglement allows us to perform DINO-style augmentation alignment on the semantic tokens while maintaining the precise spatial mapping in the localization matrix necessary for pixel-level reconstruction. We demonstrate that as few as 16 sparse tokens under this factorized form are sufficient to simultaneously support high-quality reconstruction (2.60 FID) and match the semantic performance of dense backbones (79.10% ImageNet accuracy). Our results highlight STELLAR as a versatile sparse representation that bridges the gap between discriminative and generative vision by strategically separating semantic identity from spatial geometry. Code available at https://aka.ms/stellar.

1. Introduction

Learning visual representations has been a central pursuit in computer vision since the advent of deep learning (Bengio et al., 2013). Modern vision models encode raw pixels into latent features powering nearly all downstream applications. Despite advances from early convolutional networks (Lecun et al., 1998) to ResNets (He et al., 2016) and vision transformers (ViTs) (Dosovitskiy et al., 2020), the geometric format of visual representation has remained largely unchanged: a dense 2D grid of high-dimensional features, where each vector is tied to a local patch. This design is intuitive, as it mirrors the grid-like arrangement of pixels.

On the other hand, the field faces a longstanding dilemma: the pursuit of a unified, *holistic* representation that excels at both high-level semantic understanding and low-level reconstruction. While this synthesis has succeeded in natural language processing, where reconstruction tasks like bidirectional masking (BERT (Devlin et al., 2019)) or autoregressive modeling (GPT (Brown et al., 2020)) naturally induce superior semantics, it doesn't directly transfer to the vision domain. Representations learned primarily through image reconstruction (e.g., MAE (He et al., 2022), SimMIM (Xie et al., 2022)) often yield semantics that trail behind contemporary state-of-the-art methods. Consequently, recent self-supervised learning (SSL) approaches have largely diverged into two camps: those prioritizing pixel-level grounding via reconstruction, and those prioritizing rich semantics via joint-embedding invariance (Van Assel et al., 2025).

We argue that this divergence stems from an *Invariance Paradox* inherent to the dense grid format. For a dense representation to faithfully reconstruct an image, it must preserve precise spatial information which are inherently *equivariant* to transformations like cropping or shifting. Conversely, high-level semantics are *invariant* to such transformations. Traditional SSL methods such as DINO (Caron et al., 2021) attempt to force invariance onto global representations obtained from these dense grids. This creates a fundamental conflict: the model is pressured to discard spatial variance to achieve semantic alignment, yet the dense grid format requires equivariance for spatial grounding.

In this work, we show that this paradox is not an inevitable trade-off, but a byproduct of the dense representation itself. By moving away from the 2D grid towards a *sparse, factorized latent representation*, we can jointly achieve high-fidelity reconstruction and rich semantics. Our key insight is that the information necessary to describe a scene can be disentangled into two complementary, sparse factors:

1. **The "What"**: A set of sparse latent tokens representing

[1]Microsoft [2]xAI [3]Work done at Microsoft. Correspondence to: Theodore Zhengde Zhao <theodorezhao@microsoft.com>, Mu Wei <muwei@microsoft.com>.

Preprint. February 2, 2026.

invariant visual concepts.

2. **The "Where"**: A set of equivariant coefficients representing their spatial locations.

By disentangling these factors through a low-rank matrix factorization form, we enable a "semantic triage": the model is forced to reconstruct the entire image using a highly compressed bottleneck. This encourages the model to ignore stochastically redundant background pixels and focus on semantically-rich object regions. We propose **STELLAR**, a framework that achieves high-quality reconstruction from as few as 16 tokens while encoding fine-grained semantics in a fully self-supervised manner.

Our contributions are summarized as follows:

- **Sparse Representation**: We propose STELLAR, an efficient form of vision modeling that factorizes an image into a handful of sparse tokens by disentangling *what* concepts are present from *where* they are located.
- **SSL Method**: We introduce a training scheme to learn these representations without annotation. By aligning visual concepts across views using optimal transport, we enforce invariance in the "what" factor while adapting the "where" factor, inducing rich semantics.
- **Empirical Observations**: (i) STELLAR achieves a state-of-the-art balance of semantics (IN-1K linear acc. 79.10%) and reconstruction (FID 2.60), outperforming prior approaches. (ii) Our sparse image modeling induces fine-grained, region-aware semantics even without explicit dense supervision, outperforming prior work with similar training budget.

2. Related Work

Self-supervised Learning. Modern SSL generally falls into two paradigms. *Joint Embedding* (JE) methods, such as the MoCo (He et al., 2020) and DINO (Oquab et al., 2023) families, prioritize global invariance via multi-view alignment, yielding strong semantics but often losing spatial grounding. Conversely, *Masked Image Modeling* (MIM), exemplified by MAE (He et al., 2022) and SimMIM (Xie et al., 2022), emphasizes spatial equivariance through pixel reconstruction. While hybrids like iBOT (Zhou et al., 2021) and DINOv2 (Oquab et al., 2023) attempt to combine these objectives, they still rely heavily on global invariance and forgo pixel reconstruction.

Sparse Representation. A growing body of work replaces dense feature maps with compact embeddings. Sparse R-CNN (Sun et al., 2021) and Mask2Former (Cheng et al., 2022) utilize sparse queries for supervised tasks, while BLIP-2 (Li et al., 2023) and TiTok (Yu et al., 2024) employ sparse tokens for vision–language or generative efficiency. SemMAE (Li et al., 2022) utilizes sparse tokens to guide masking using a pretrained teacher. Unlike these methods, STELLAR treats sparse tokens as the **primary latent representation** and learns in SSL manner.

Disentanglement & Low-rank Factorization. The assumption that high-dimensional data lie on low-dimensional manifolds is foundational to dictionary learning (Mairal et al., 2008). In deep learning, low-rank constraints are typically applied to weights for efficiency (e.g., LoRA (Hu et al., 2022)). STELLAR differs by applying *low-rank factorization to the feature map itself*, disentangling "what" (semantic latents $\boldsymbol{L}$) from "where" (spatial assignments $\boldsymbol{S}$).

The Empirical Dilemma. Current vision frameworks face a persistent gap: models excelling at pixel-level reconstruction often produce weaker semantic representations (Zhang et al., 2022; Chen et al., 2024), while those achieving top-tier semantics often abandon reconstruction to avoid low-level shortcuts (Assran et al., 2023; Darcet et al., 2025). We demonstrate that by factorizing the latent representation, it is possible to achieve strong performance on both image understanding and reconstruction.

3. Preliminaries

Representation learning involves encoding an image $X \in \mathcal{X}$ to latent features $\boldsymbol{Z}(X)$ for downstream tasks. Traditionally, vision representations take a *dense* spatial form:

$$\boldsymbol{Z} \in \mathbb{R}^{n \times d},$$

where $n = h \times w$ denotes the number of patches on a dense grid that partitions the image. Each grid location is represented by a feature vector $\mathbf{z}_i := \boldsymbol{Z}_{i,:} \in \mathbb{R}^d$ for $1 \leq i \leq n$. Most vision architectures also incorporate a global representation $\mathbf{z}_0 \in \mathbb{R}^d$, typically obtained via global pooling or a specialized `[CLS]` token that undergoes self-attention with patch tokens.

Ideally, we want $\boldsymbol{Z}$ to serve as a *holistic* representation of the image X, which retains sufficient information about the image details, while at the same time possesses rich semantics for downstream tasks. Mathematically, we define such representation as follows:

- **Reconstruction**: There exists a decoder $\mathcal{D}$ such that $\mathcal{D}(\mathbf{Z}(X)) \approx X$. This ensures the representation is spatially and texturally grounded in the physical input.
- **Semantics**: For a downstream task with joint distribution $(X, Y) \sim \mathcal{X} \times \mathcal{Y}$, there exists a simple predictor $f \in \mathcal{F}$ (e.g., a linear layer) such that the expected task loss $\mathbb{E}_{(X,Y)}\big[\mathcal{L}(f(\mathbf{Z}(X)), Y)\big]$ is minimized using frozen features. Typically Y reflects human perception.

Current SSL paradigms are caught in a fundamental "*Invariance Paradox*". In order to learn high-level semantics, Joint Embedding (JE) methods (e.g. the DINO family) impose *invariance* to spatial transformations, even when the image is cropped to as small as only 5%. On the other hand, reconstruction requires spatial detail, because every pixel shift requires a different set of features for precise reconstruction. This results in representations which are highly *equivariant* to the transformation, i.e. the feature map transforms along with the transformation in the image.

Let $\mathcal{T}$ be a group of spatial transformations (e.g., translations), and $t_\theta \in \mathcal{T}$ is parametrized by θ i. A representation $\boldsymbol{Z}(X)$ suffers from the *Invariance Paradox* if it must simultaneously satisfy two contradictory constraints:

- **Semantic Invariance:** The representation should be insensitive to $t_\theta \in \mathcal{T}$:
$$\|\frac{\partial}{\partial\theta}\boldsymbol{Z}(t_\theta \circ X)\|_F \approx 0.$$
- **Spatial Equivariance:** To allow for high-fidelity reconstruction, the representation must track spatial shifts: $\mathcal{D}(\boldsymbol{Z}(t_\theta \circ X)) \approx t_\theta \circ X$. With chain rule and matrix norm inequalities, we have
$$\|\frac{\partial}{\partial\theta}\boldsymbol{Z}(t_\theta \circ X)\|_F \gtrsim \frac{\left\|\frac{\partial(t_\theta \circ X)}{\partial\theta}\right\|_F}{\sigma_{\max}\left(\frac{\partial\mathcal{D}}{\partial\boldsymbol{Z}}\right)} > 0.$$

4. The STELLAR Framework

4.1. Sparse Image Modeling

From now on we consider a form of representation in alternative to the dense grid-based representations describing what appears at each individual location. We start from the principle that an image depicts the physical world, which can be understood as a collection of objects located in space.

To begin with, we model an image with a compact set of semantic concepts together with their spatial distributions. Let there be r concept embeddings $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_r \in \mathbb{R}^d$, where each $\boldsymbol{s}_j$ captures a distinct semantic concept. The spatial distribution of these concepts is expressed through weights $\boldsymbol{l}_1, \cdots, \boldsymbol{l}_n \in \mathbb{R}^r$, where n is the total number of patches.

By constraining $0 \leq \boldsymbol{l}_i \leq 1$ and $\mathbf{1}^\top \boldsymbol{l}_i = 1$, each patch is represented as a convex combination of the concept embeddings: $\boldsymbol{v}_i = \sum_{j=1}^r \boldsymbol{l}_{i,j}\boldsymbol{s}_j$. Thus, the set $\boldsymbol{s}_j{}_{j=1}^r$ acts as a basis for constructing local features. In matrix form, the latent representation now takes the form

$$\boldsymbol{Z}(X) = \boldsymbol{L}(X)\boldsymbol{S}(X), \tag{1}$$

where $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r]^\top \in \mathbb{R}^{r\times d}$ is the *semantic matrix*, and $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_n]^\top \in \mathbb{R}^{n\times r}$ is the *localization matrix*, with the constraint $0 \leq \boldsymbol{L} \leq 1, \boldsymbol{L}\mathbf{1}_r = \mathbf{1}_n$.

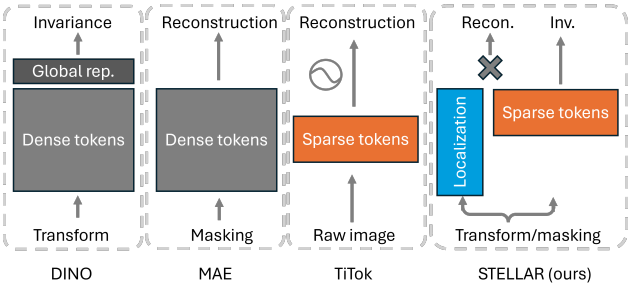

Figure 1. Comparison of learning different latent representation.

Compared to a canonical dense representation of shape $n\times d$, $\boldsymbol{Z} = \boldsymbol{L}\boldsymbol{S}$ can be considered as a form of low-rank matrix approximation from the sparse representation. While the form resembles the low-rank structure used in convex semi-nonnegative matrix factorization (Ding et al., 2008), $\boldsymbol{S}$ and $\boldsymbol{L}$ are not obtained from any matrix factorization algorithm, but are instead direct output from the forward pass of the encoder, allowing end-to-end training using SSL objectives.

4.2. Equivariant Partitioning

The factorized form in equation 1 not only provides a more efficient latent representation ($r(n+d) \ll nd$. With 16 tokens, ViT-base on a 224×224 image enjoys 90% reduction), it also provides an escape from the invariance paradox. The spatial transformation is now partitioned as follows:

$$\underbrace{\frac{\partial\boldsymbol{Z}(t_\theta \circ X)}{\partial\theta}}_{\text{Total Equivariance}} = \underbrace{\left(\frac{\partial\boldsymbol{L}(t_\theta \circ X)}{\partial\theta}\right)\boldsymbol{S}}_{\text{Spatial Equivariance}} + \underbrace{\boldsymbol{L}\left(\frac{\partial\boldsymbol{S}(t_\theta \circ X)}{\partial\theta}\right)}_{\text{Semantic Variance}\approx 0}. \tag{2}$$

With the spatial and semantic information disentangled, we can offload the spatial equivariance entirely to localization matrix $\boldsymbol{L}$, while still achieving semantic invariance in $\boldsymbol{S}$. We illustrate the learning paradigm of the factorized representation in Fig. 1. We require that $\boldsymbol{Z}(X)$ can reconstruct the image by minimizing

$$\mathcal{L}_{recon} = \ell(\mathcal{D}(\boldsymbol{L}(X)\boldsymbol{S}(X)), X). \tag{3}$$

This *low-rank approximated reconstruction* forces the model to use only sparse tokens $\{\boldsymbol{s}_j\}_{j=1}^r$ to capture sufficient information about the image.

4.3. Vision Concept Clustering

To encourage sparse tokens to represent transferable vision concepts, we structure them into K learnable prototypes $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K \in \mathbb{R}^p$. A backbone encoder $\mathcal{E}$ maps a mini-batch of m images into sparse features $\boldsymbol{S}^1, \cdots, \boldsymbol{S}^m$. Each token is projected onto the unit sphere $\mathbb{S}^{p-1}$ via a normalized

projector $h : \mathbb{R}^d \to \mathbb{S}^{p-1}$, and its similarity to prototypes $\boldsymbol{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K]$ gives logits

$$\lambda_j^i = [\boldsymbol{c}_1 \cdot h(\boldsymbol{s}_j^i), \cdots, \boldsymbol{c}_K \cdot h(\boldsymbol{s}_j^i)], \quad j = 1, \ldots, r. \tag{4}$$

Soft assignments over the prototypes is obtained with

$$q_{j,k}^i = \frac{\exp(\lambda_{j,k}^i/\tau)}{\sum_{k'=1}^K \exp(\lambda_{j,k'}^i/\tau)}, \tag{5}$$

where τ controls sharpness. Direct entropy minimization of q_j^i is unstable due to non-convexity and empty clusters. Following (Caron et al., 2020; Darcet et al., 2025), we compute balanced assignments $\tilde{q}_j^i$ from q_j^i using the Sinkhorn-Knopp algorithm (see appendix) without gradient, and minimize

$$\mathcal{L}_{\text{cluster}} = -\frac{1}{mr} \sum_{i=1}^m \sum_{j=1}^r \sum_{k=1}^K \tilde{q}_{j,k}^i \log q_{j,k}^i. \tag{6}$$

Unlike DINOv2 and SwAV which only use Sinkhorn for balancing teacher targets, we explicitly minimize $\mathcal{L}_{\text{cluster}}$ along with all other objectives.

4.4. Set Concepts Alignment

To achieve the semantic invariance in equation 2, we align the sparse tokens $\boldsymbol{s}'_1, \ldots, \boldsymbol{s}'_r$ obtained from a transformed view (e.g. masking or cropping) to the ones from the global view $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r$. However, this set concepts alignment problem is challenging compared to global representation alignment in traditional JE methods, because there is no inherent ordering in the r tokens. To solve the problem, we apply optimal transport with the cost matrix

$$\Theta_{j'j} = \|\boldsymbol{s}'_{j'} - \boldsymbol{s}_j\|_2. \tag{7}$$

We solve for an assignment matrix $\boldsymbol{P}$ via entropy-regularized optimal transport:

$$\min_{\boldsymbol{P} \geq 0} \quad \sum_{j',j} \boldsymbol{P}_{j'j} \Theta_{j'j} - \epsilon H(\boldsymbol{P}), \tag{8}$$

$$s.t. \quad \boldsymbol{P}\mathbf{1}_r = \boldsymbol{P}^T \mathbf{1}_r = \frac{1}{r}\mathbf{1}_r, \tag{9}$$

with $H(\boldsymbol{P}) = -\sum j', j \boldsymbol{P}_{j'j} \log \boldsymbol{P}_{j'j}$. We solve for $\boldsymbol{P}$ using the Sinkhorn algorithm, and define the matching $\sigma(j') := \text{argmax}_j \boldsymbol{P}_{j'j}$. Compared to bipartite matching algorithms such Hungarian matching widely used in previous literature, this algorithm is up to $100\times$ faster, with experimental results analyzed in the appendix.

We then compute prototype assignments for the transformed view tokens $q'_{j'} = \text{softmax}(\boldsymbol{C}^T h(\boldsymbol{s}'_{j'})/\tau)$, and minimize the set concept alignment loss

$$\mathcal{L}_{\text{align}} = -\frac{1}{r} \sum_{j'=1}^r \sum_{k=1}^K \tilde{q}_{\sigma(j'),k} \log q'_{j',k}. \tag{10}$$

Optionally, we use the same framework to cluster and align the CLS token with its own projector and prototypes, similar to previous JE methods. However, we do not used it for reconstruction. We also apply KoLeo regularization (Sablayrolles et al., 2018) on the normalized sparse tokens $\bar{\boldsymbol{s}}_j := \boldsymbol{s}_j/\|\boldsymbol{s}_j\|$ obtained from the same image to encourage concept diversification:

$$\mathcal{L}_{KoLeo} = -\frac{1}{r} \sum_{j=1}^r \log \left(\min_{j' \neq j} \frac{1}{2} \|\bar{\boldsymbol{s}}_j - \bar{\boldsymbol{s}}_{j'}\|_2 \right). \tag{11}$$

All together, we jointly optimize the following objectives by training the encoder $\mathcal{E}$, decoder $\mathcal{D}$, projector h, and prototypes $\boldsymbol{C}$ jointly with the final objective:

$$\min_{\mathcal{E},\mathcal{D},h,\boldsymbol{C}} \quad a_1 \mathcal{L}_{\text{recon}} + a_2 \mathcal{L}_{\text{cluster}} + a_3 \mathcal{L}_{\text{align}} + \tag{12}$$

$$a_4 \mathcal{L}_{\text{cluster-cls}} + a_5 \mathcal{L}_{\text{align-cls}} + a_6 \mathcal{L}_{\text{KoLeo}}. \tag{13}$$

In summary, we proposed a sparse vision representation $(\boldsymbol{S}, \boldsymbol{L}) = \mathcal{E}(X)$ that explicitly disentangles semantic concepts from their spatial distributions, enabling the latent variables to support both pixel-level reconstruction and high-level semantic understanding. We introduced a simple encoder design to obtain these latent variables and SSL objectives to shape them into transferable visual concepts.

We refer to our framework of learning the spatial-semantic factorized representation $\boldsymbol{Z}(X) = \boldsymbol{L}(X)\boldsymbol{S}(X)$ as **S**parse **T**oken **E**xtraction and **L**ocalization with **L**ow-rank **A**pproximated **R**econstruction (STELLAR).

4.5. Model Design

We note that the framework only specifies the latent space, and does not prescribe any specific encoder or decoder architecture. In this work, we adopt a simple design with common modules and model architectures to obtain $\boldsymbol{S}$ and $\boldsymbol{L}$ as described below.

For the encoder part, we use an existing ViT(Dosovitskiy et al., 2020) as the backbone, and equip it with r learnable latent query vectors, which are passed to the transformer blocks alongside the patch tokens. Processed by the ViT jointly, the latent queries produce sparse tokens $\boldsymbol{S} \in \mathbb{R}^{r \times d}$.

To obtain the localization matrix $\boldsymbol{L} \in \mathbb{R}^{n \times r}$ associated with the sparse tokens, we use the dense feature map $\boldsymbol{U} \in \mathbb{R}^{n \times d}$ output from the image patches. We project both $\boldsymbol{S}$ and $\boldsymbol{U}$ into a shared embedding space and compute their pairwise cosine similarities, followed by a softmax normalization with temperature τ_{spatial} along the the second dimension:

$$\boldsymbol{L} = \text{softmax}\left(\text{cossim}(\boldsymbol{U}\boldsymbol{W}_1, \boldsymbol{S}\boldsymbol{W}_2)/\tau_{\text{spatial}}\right). \tag{14}$$

$\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are learnable linear projections, and τ_{spatial} controls the sharpness of the spatial distribution. We note that

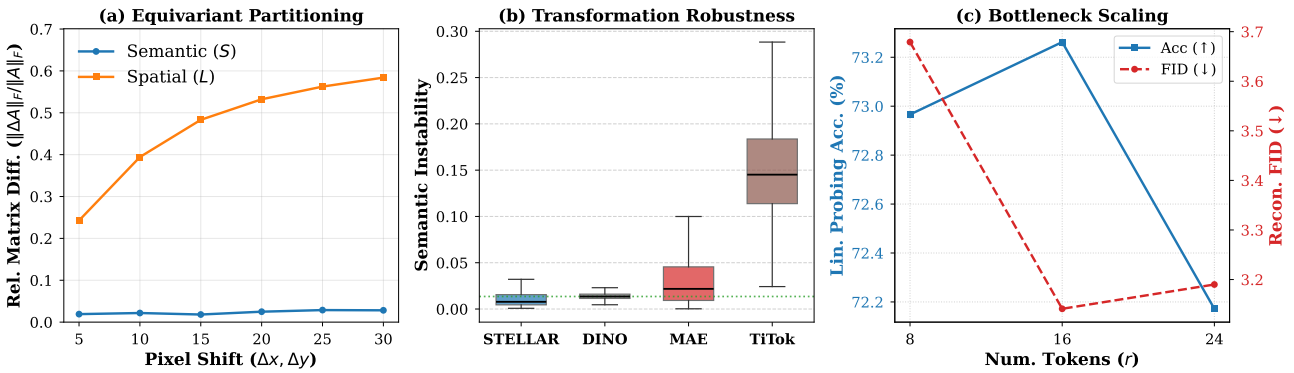

Figure 2. Analysis of STELLAR representation. (a) Relative matrix difference in $\boldsymbol{L}$ and $\boldsymbol{S}$ under controled pixel shift in the input image. (b) Cosine distance of latent representation under random 50-100% random cropping. (c) Impact of number of sparse tokens r on reconstruciton and semantic quality.

his mapping is structurally similar to the attention weights obtained in a single-head cross-attention layer, up to the use of L2 normalization and an explicit temperature parameter. Therefore, the latent representation $\boldsymbol{Z} = \boldsymbol{LS}$ can be viewed as rebuilding a dense feature map for reconstruction by cross-attending to only r sparse concept tokens.

All together, the encoder $\mathcal{E}$ includes ViT transformer blocks, r learnable latent query vectors, and projection layers $\boldsymbol{W}_1, \boldsymbol{W}_2$. The decoder $\mathcal{D}$ is a 6-layer lightweight ViT reconstructing the image patches.

The STELLAR framework can be used on a pretrained ViT such as MAE or DINO to leverage the foundation prior and shape it into a sparse holistic representation. It can also be trained from a random prior and reach competitive spatial, semantic, and reconstruction quality. We provide deep analysis in the ablation study.

5. Experiments

We train STELLAR on ImageNet-1K (Deng et al., 2009) without labels. The encoder is a vanilla ViT (Dosovitskiy et al., 2020) augmented with 8–24 learnable latent queries that produce sparse tokens. A lightweight 6-layer ViT serves as the decoder predicting either MaskGIT-VQGAN tokens (Esser et al., 2021; Chang et al., 2022). When using a foundation prior for the backbone ViT, we ensure that the pretraining was also performed only on ImageNet-1K. We use MAE as the default prior and studied the effect of different prior modes in ablation study. When training from random prior, we use a momentum updated encoder to encode the target assignments $\tilde{q}$ in equation 6 and equation 10, following Grill et al.; Caron et al..

5.1. Probing the Factorized Representation

We designed a series of experiments to analyzed the factorized representation $\boldsymbol{Z} = \boldsymbol{LS}$ from STELLAR training.

Experiment 1: Equivariant Partitioning We built a controllable and parametrized spatial transformation group to examine the equivariant partitioning in equation 2. Given an image, we take a crop and shift it gradually from 5 to 30 pixel, either horizontally or vertically. We calculated the relative matrix difference $\frac{\|\boldsymbol{S}(t_\theta \circ X) - \boldsymbol{S}(X)\|_F}{\|\boldsymbol{S}(X)\|_F}$ and $\frac{\|\boldsymbol{L}(t_\theta \circ X) - \boldsymbol{L}(X)\|_F}{\|\boldsymbol{L}(X)\|_F}$. Optimal tsansport in equation 8 is used to match the token ordering. As shown in Fig. 2(a), the semantic matrix $\boldsymbol{S}$ stays almost completely invariant, while the spatial localization matrix $\boldsymbol{L}$ changes continuously with the spatial shift, proving the effectiveness of equivariant partitioning in the factorized representation.

Experiment 2: Transformation Robustness Next we compare the semantic stability of STELLAR representation with baseline models under random resized cropping at scale 50-100%. We calculate the cosine distance from the feature of the untransformed image as a measure of semantic instability. For DINO and MAE, we used mean-pooled dense features, and for sparse models STELLAR and TiTok, we used mean-pooled sparse tokens. As shown in Fig. 2(b), the sparse tokens of STELLAR enjoys high transformation robustness at DINO level. As expected, the reconstruction-based models MAE and TiTok show higher variance to spatial transformation. Specifically, the un-factorized sparse representation from TiTok is extremely unstable, as the model need to store both semantic and spatial information in the same sparse tokens for reconstruction.

Experiment 3: Effect of Low-rank Bottleneck The number of sparse tokens r serves as the intrinsic rank of the latent representation. We experimented scaling r from 8 to 24, and evaluated reconstruction with FID (Heusel et al., 2017) and semantics with linear probing on mean-pooled sparse tokens. As shown in Fig. 2(c), the linear probing accuracy decreases as r increases, while reconstruction improves with more tokens, showing a trade-off in the intrinsic rank of the

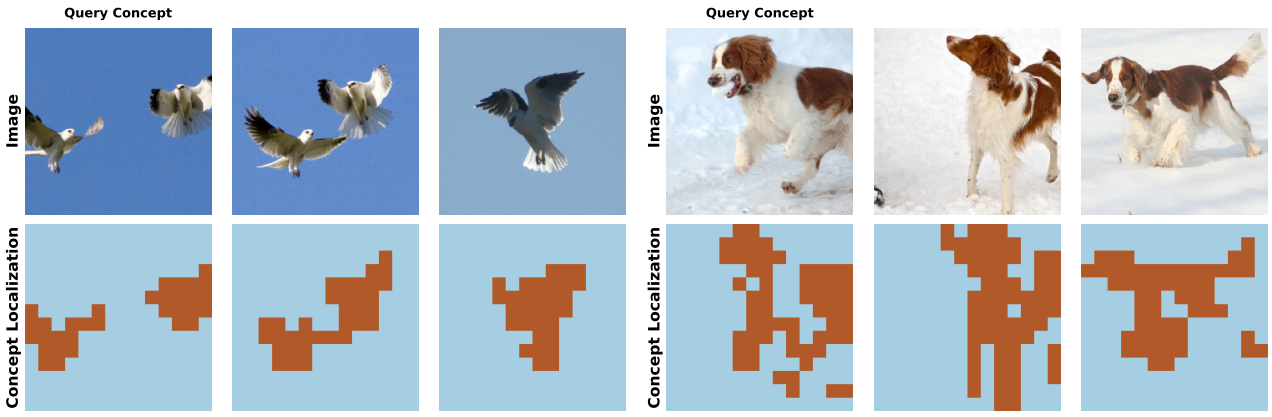

Figure 3. Vision concepts retrieved from the training set. We show show both the image and the spatial localization of the concept.

representation. As a sweet spot, $r = 16$ enjoys both rich semantics and high-quality reconstruction, which we used as the default for all other experiments.

Finally, we visualize the factorized representation in Fig. 3. We show the spatial localization (thresholded by $1/r$) of a sparse token in the image, and the top retrieved semantic concepts from the training dataset.

5.2. Evaluating Holistic Representation

Next we examine the semantic quality and reconstruction potential of representation from different models. We trained the same decoder in STELLAR on top of the frozen features from different encoders. The original decoder in STELLAR was also finetuned with the rest of the model frozen. TiTok used it's native decoder, which is a full-sized ViT compared to only 6 layers in STELLAR.

As shown in Table 1, STELLAR (shown as ours) shows superior performance in supporting both semantics and reconstruction. The linear probing and k-NN accuracy of STELLAR surpass reconstruction-feasible representations from all baselines, despite trailing behind the CLS token from DINO, which is infeasible as a reconstruction latent.

On reconstruction, STELLAR shows comparable FID and LPIPS loss (Zhang et al., 2018) to the dense feature map from MAE, with 90% reduction in latent size. Although TiTok achieved lower FID with a much larger decoder, its shows highest LPIPS loss, indicating poor spatial consistency. In contrast, STELLAR exhibits superior reconstruction locality even with fewer tokens. The full-rank dense feature map $\boldsymbol{U}$ (196 tokens) from the ViT in STELLAR shows even lower reconstruction FID, while drops in semantic quality. Finally, when scaling to huge-sized ViT, STELLAR achieves top reconstruction and semantic quality, even without decoder finetuning.

Table 1. Reconstruction and semantic metrics on IN1K of STELLAR (ours) and baseline models. For reference, we also reported semantic metrics of the global representation from DINO, and huge size STELLAR model. Best main results are shown in **bold**. Model sizes are ViT-B by default, with larger sizes indicated in parentheses. *: TiTok used its native ViT decoder of larger size.

		RECONSTRUCTION		SEMANTICS	
MODEL	# TKS	FID ↓	LPIPS ↓	LIN.	KNN
DINO	1	-	-	76.46	74.69
DINO	196	3.27	0.2121	70.31	54.41
MAE	196	3.02	**0.2071**	66.32	25.82
TITOK*	32	2.75	0.3281	33.42	7.30
TITOK*	64	1.99	0.2571	32.87	7.29
OURS	16	3.06	0.2077	**73.26**	**67.25**
OURS	196	**2.85**	0.2085	72.21	64.71
OURS(H)	16	2.60	0.1729	79.10	77.31

5.3. Benchmarking Image Understanding

Lastly, we benchmark STELLAR in classical image understanding tasks with linear probing on frozen features, comparing against other ImageNet-pretrained SSL models. We report results for classification on ImageNet-1K (IN1K), Oxford-IIIT Pet (Pets) (Parkhi et al., 2012), Food-101 (Food) (Bossard et al., 2014), and GlaS (Sirinukunwattana et al., 2016) for cancer grade classification in histopathology. Segmentation benchmarks include ADE20K (Zhou et al., 2017b), Cityscapes (Cordts et al., 2016), and Pascal VOC (Everingham et al., 2010). For broader context, we also include AIM (El-Nouby et al., 2024) and DINOv2 (Oquab et al., 2023), which leverage substantially larger training corpora (100–1000× more images).

As shown in Table 2, the feature map from STELLAR achieves superior performance on ADE20K and Pascal VOC, showing strong fine-grained understanding despite not applying SSL objectives directly to the dense feature map $\boldsymbol{U}$. Sparse token modeling implicitly organizes the

Table 2. **Evaluation of Fine-grained and Global Image Understanding.** We evaluate semantic segmentation (mIoU %) and classification accuracy (%) via linear probing on frozen features. We used the dense feature map from the backbone for all segmentation tasks and all models. **Bold**: best with ImageNet training. Underline: best in architectural class (e.g., ViT-B).

		SSL Type		Segmentation (mIoU)			Classification (Acc)			
Model	**Arch.**	**Target**	**Method**	**ADE20K**	**CitySc**	**VOC**	**IN1K**	**Pets**	**Food**	**GlaS**
Semantic-Centric (Joint Embedding / Invariance)										
BYOL	RN-50	GLOBAL	DISTILL	18.43	18.66	63.89	70.39	82.77	64.57	95.00
MOCO V3	VIT-B	GLOBAL	CONTR.	29.45	25.13	74.08	74.31	91.14	77.47	**97.50**
DINO	VIT-B	GLOBAL	DISTILL	26.87	26.82	79.29	76.46	**93.84**	79.28	95.00
MSN	VIT-B	GLOBAL	MASKING	26.66	25.39	68.59	73.65	75.91	68.93	92.50
DENSECL	RN-50	DENSE	CONTR.	23.08	18.63	70.95	61.10	72.99	59.16	85.00
DATA2VEC	VIT-B	DENSE	LAT-MIM	22.03	23.49	61.33	54.90	26.47	34.40	73.75
SIAMESEIM	VIT-B	DENSE	LAT-MIM	29.24	26.52	81.38	74.97	91.61	71.01	91.25
I-JEPA	VIT-H	DENSE	LAT-MIM	21.57	18.59	74.13	71.72	84.68	70.34	87.50
IBOT	VIT-B	GL+DE	DIST+MIM	31.78	25.69	77.06	76.40	92.40	78.08	96.25
IBOT	VIT-L	GL+DE	DIST+MIM	33.26	26.37	77.57	78.53	92.12	**81.07**	96.25
Image-Centric (Reconstruction)										
BEIT	VIT-B	DENSE	TOK MIM	11.58	18.90	27.44	32.94	36.20	54.49	90.00
BEIT	VIT-L	DENSE	TOK MIM	12.64	20.37	25.48	36.77	36.71	56.03	90.00
SIMMIM	SWIN-B	DENSE	PIX MIM	12.46	17.23	35.14	24.77	27.39	40.94	77.50
MAE	VIT-B	DENSE	PIX MIM	30.91	29.44	76.43	66.32	81.58	70.40	93.75
MAE	VIT-L	DENSE	PIX MIM	34.36	32.53	77.79	73.09	84.30	76.22	95.00
MAE	VIT-H	DENSE	PIX MIM	36.16	**35.21**	78.07	75.22	84.96	78.36	95.00
SEMMAE	VIT-B	DENSE	PIX MIM	3.52	25.48	48.33	43.84	56.99	58.90	92.50
TITOK-64	VIT-B	SPARSE	SPRS REC	–	–	–	32.87	42.06	43.68	**97.50**
TITOK-32	VIT-L	SPARSE	SPRS REC	–	–	–	33.42	27.83	38.83	78.75
Our Method (Sparse Factorized Modeling)										
STELLAR	VIT-B	SPARSE	INV+REC	31.33	27.74	81.83	73.26	89.70	74.09	95.00
STELLAR	VIT-L	SPARSE	INV+REC	34.02	31.32	**85.90**	76.94	92.53	74.78	**97.50**
STELLAR	VIT-H	SPARSE	INV+REC	**36.66**	33.30	85.66	**79.10**	92.53	77.43	92.50
Larger Scale Pretraining Beyond ImageNet (Reference Only)										
AIM	600 M	DENSE	IMAGE AR	29.00	27.04	64.55	63.78	64.68	75.19	98.75
AIM	1 B	DENSE	IMAGE AR	29.59	27.05	63.90	66.86	64.21	77.96	96.25
DINOV2	VIT-B*	GL+DE	DIST+MIM	40.10	34.66	89.52	82.82	95.59	91.08	98.75
DINOV2	VIT-L*	GL+DE	DIST+MIM	40.45	32.07	89.19	84.23	96.08	92.94	98.75

feature map into semantic regions: to reconstruct the image, each token must encode information covering all spatial parts of the scene, resulting in region-aware representations. While MAE leads in CityScapes, STELLAR follows closed with performance comparable to MAE and DINOv2.

On global image understanding tasks, STELLAR achieves the highest accuracy on IN1K at large model scale, but smaller variants underperform methods such as DINO, which explicitly optimizes for global representations. In general, STELLAR outperforms image reconstruction models and most JE methods, but trails behind top JE models in global semantics. As we do not model the image as a single concept, averaging token features can dilute discriminative information, which is particularly detrimental on object-centric datasets like Pets and Food. Interestingly, on histopathology images involving complex tissue microenvironments, STELLAR achieves the best performance. These results indicate that STELLAR excels at modeling complex, multi-object scenes, while global classification on simple object-centric datasets remains more challenging.

5.4. Ablation Analysis

Low-rank approximated reconstruction. As shown in Table 3, removing the low-rank reconstruction objective (A) reduces both global and fine-grained understanding. Since the remaining objectives resemble typical SSL methods, the model still retains reasonable global performance, but fine-grained understanding suffers more. This indicates that low-rank reconstruction encourages sparse tokens to serve as holistic representations covering the entire image.

Concept clustering. Eliminating online clustering and set alignment (B) leads to a sharp drop in understanding, highlighting the necessity of structuring sparse tokens into view-invariant concepts. Even when the alignment loss is present (D), missing the clustering loss still lead to feature collapse.

Set alignment. The training collapsed when training with only reconstruction and clustering (C), underscoring the critical role of set concepts alignment. Additional alignment on the CLS token (E) primarily benefits global classification

Table 3. **Ablation.** We isolate the impact of each objective on semantic abstraction (IN1K) and spatial grounding (ADE20K), and reconstruction (FID). *Default* denotes the full STELLAR framework. All results are based on ViT-B.

	Recon.	Cluster	Set Align	CLS Align	KoLeo	rFID ↓	IN1K ↑	ADE ↑
DEFAULT	✓	✓	✓	✓	✓	**3.14**	**73.26**	**31.33**
Impact of Individual Components								
(A)	✗	✓	✓	✓	✓	—	72.44 (-0.82)	29.94 (-1.39)
(B)	✓	✗	✗	✗	✓	3.21 (+0.07)	52.07 (-21.19)	20.46 (-10.87)
(C)	✓	✓	✗	✗	✓	8.95 (+5.81)	2.73 (-70.53)	1.93 (-29.39)
(D)	✓	✗	✓	✓	✓	3.62 (+0.48)	42.14 (-31.12)	18.90 (-12.43)
(E)	✓	✓	✓	✗	✓	3.26 (+0.12)	70.79 (-2.47)	30.20 (-1.12)
(F)	✓	✓	✓	✓	✗	3.25 (+0.11)	72.05 (-1.21)	30.10 (-1.23)

but has limited effect on spatial grounding. Finally, KoLeo regularization (F) consistently improves all tasks at similar level. Interestingly, the absent of either concept clustering or set alignment led to a sharp drop in performance.

Foundational Prior We ablated STELLAR trained from different pretrained foundational prior in Table 4. We observes that STELLAR significantly boost the semantic quality from MAE, and the spatial grounding from DINO prior. The semantics performance falls to similar level despite different foundational priors. When training from random prior, STELLAR is able to reach semantics at MAE level and spatial understanding similar to that from DINO prior. The reconstruction quality stays consistent in all cases.

Table 4. Evaluating STELLAR trained from different foundational priors. *Base* represents the performance of the original backbone.

Prior	Recon FID ↓	Semantic (IN1K) BASE	Semantic (IN1K) +STELLAR	Spatial (ADE20K) BASE	Spatial (ADE20K) +STELLAR
MAE	**3.14**	66.32	**73.26** (+6.9)	30.91	**31.33** (+0.4)
DINO	3.31	76.46	73.31 (-3.2)	26.87	**28.17** (+1.3)
RAND	3.21	–	65.28	–	28.10

6. Discussion and Concluding Remarks

We have demonstrated that the long-standing trade-off between *semantic abstraction* and *spatial grounding* (the Invariance Paradox) is largely an artifact of the traditional dense-grid representation. By factorizing the latent space into sparse "What" and "Where" components, STELLAR effectively resolves this conflict.

The core of STELLAR's success lies in the principle of *semantic triage*. In a dense model (e.g., MAE), the representation is spatially exhaustive but semantically diluted; the model is forced to allocate representational capacity to every patch, including stochastically redundant background noise.

The low-rank factorization provides the mathematical machinery to disentangle two distinct types of visual information. The *Concept Tokens* ($\mathbf{S}$) are trained to be view-invariant, capturing the categorical "What," while the *Spatial Coefficients* ($\mathbf{L}$) remain equivariant, capturing the geometric "Where."

This disentanglement allows the model to satisfy the Joint Embedding objective (alignment across views) without destroying the spatial anchors needed for high-fidelity reconstruction. By separating these factors, we avoid the "semantic blurring" often seen in global-pooling methods, as each sparse token maintains a precise, albeit flexible, relationship with the physical image geometry.

The Path to Unified Multimodality: Perhaps the most promising frontier for STELLAR is its potential as a visual front-end for Large Language Models (LLMs). Because our tokens are sparse and semantically grounded, they offer a more natural interface for cross-modal alignment than the hundreds of dense tokens generated by standard ViTs. Future work will investigate the systematic integration of STELLAR concepts with linguistic embeddings for more interpretable multimodal understanding.

Overall, our results highlight sparse tokens as a promising direction for unifying efficiency, interpretability, and semantic richness in self-supervised representation learning.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

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

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pp. 1298–1312. PMLR, 2022.

Bao, H., Dong, L., and Wei, F. Beit: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Chen, X., Liu, Z., Xie, S., and He, K. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024.

Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875, 2021.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Darcet, T., Baldassarre, F., Oquab, M., Mairal, J., and Bojanowski, P. Cluster and predict latent patches for improved masked image modeling. *arXiv preprint arXiv:2502.08769*, 2025.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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.

Ding, C. H., Li, T., and Jordan, M. I. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

El-Nouby, A., Klein, M., Zhai, S., Bautista, M. A., Toshev, A., Shankar, V., Susskind, J. M., and Joulin, A. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2):303–338, 2010.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Li, G., Zheng, H., Liu, D., Wang, C., Su, B., and Zheng, C. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Loshchilov, I. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., and Bach, F. Supervised dictionary learning. *Advances in neural information processing systems*, 21, 2008.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sablayrolles, A., Douze, M., Schmid, C., and Jégou, H. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

Sirinukunwattana, K., Pluim, J. P. W., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D. R. J., and Rajpoot, N. M. Gland segmentation in colon histology images: The glas challenge contest, 2016.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14454–14463, 2021.

Tao, C., Zhu, X., Su, W., Huang, G., Li, B., Zhou, J., Qiao, Y., Wang, X., and Dai, J. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2132–2141, 2023.

Van Assel, H., Ibrahim, M., Biancalani, T., Regev, A., and Balestriero, R. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*, 2025.

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense contrastive learning for self-supervised visual pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., and Chen, L.-C. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024.

Zhang, L., Yang, Q., and Agrawal, A. Assessing and learning alignment of unimodal vision and language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14604–14614, 2025.

Zhang, M., Xiao, T. Z., Paige, B., and Barber, D. Improving vae-based representation learning. *arXiv preprint arXiv:2205.14539*, 2022.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017a.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017b.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A. Implementation Details

A.1. STELLAR Training

We trained STELLAR with ViT models at size base, large, and huge, along with the latent queries, projection layers, clustering head, and a 6-layer ViT decoder. In the default setting, we initialized the ViT part in the encoder from public MAE checkpoint, and trained for 150 epochs for STELLAR-B, 100 epochs for STELLAR-L, and 50 epochs for STELLAR-H. We used 16 NVIDIA A100-80GB with batch size 128 each, totaling 2048. We used AdamW(Loshchilov, 2017) with base learning rate 1.5×10^{-4} for STELLAR-B, and 5×10^{-5} for STELLAR-L and STELLAR-H.

For concept clustering, we used 16384 prototypes for sparse and CLS tokens each. The projector is a 2-layer MLP before the prototype layer. We used 3 steps of Sinkhon-Knopp algorithm. The temperature in sparse-dense cosine similarity softmax is 0.06. We used 6-8 random masked views to align the sparse tokens, and additional 6-8 local crops to align the CLS token. Global views are of random scale 36% to 100%, and local view are of random scale 6% to 36%. We also apply color jittering, grascaling and Gaussian blurring.

In the ablation study of random prior, we trained the model from scratch and used exponential moving average (EMA) updated momentum encoder to encode the target prototype assignments in the warm-up stage. We EMA updated the full encoder (ViT, latent queries, projection, clustering head with momentum 0.996. The momentum encoder was used to encode a global view of the image into target prototype assignments, for both clustering loss and alignment loss. The masking ratio was 0.6 in the warm-up stage, and 0.8 during standard training. We trained the model with 150 epochs of EMA warm-up and 75 epochs of standard training.

A.2. Evaluation Protocol

For STELLAR and all baseline models, we evaluated the frozen feature from the pretrained model with linear probing. We used layer norm in classification tasks, and batch norm in segmentation tasks, followed by a single linear layer predicting the class of the image or patch. For all benchmarks, we split 10% from the training set for validation. We tuned hyper-parameter with learning rate $1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}$, and batch size 64, 128, 256, 512, 1024, 2048, 4096, 8192.

As the SSL methods varies across different baseline models, for classification tasks we used the mean-pooled feature from the representations where the corresponding SSL method was performed, e.g. the global CLS token for DINO, and dense patch tokens for MAE. We noticed the linear probing accuracy can vary depending on the pooling choice, and conducted experiments by using different types of tokens for each model, with results in Table 5. We observed that the SSL-ed are typically the best choice for linear probing, except for iBOT, which highly relies on the global CLS token for classification, even though the model was trained with MIM. In contrast, STELLAR and MAE are relatively more robust to token choices.

Table 5. ImageNet-1K linear probing accuracy (%) by pooling different tokens. We mark in **bold** the tokens on which the specific SSL method was applied, and the top accuracy for each method.

	DINO		MAE		iBOT			STELLAR (ours)	
tokens	**global**	dense	global	**dense**	global	dense	**gl.+de.**	**sparse**	dense
lin. acc.	**76.46**	70.31	65.61	**66.32**	**76.40**	71.44	71.58	**73.26**	72.21

B. Additional Results

B.1. Effect of pretraining data

We pretrained separate STELLAR versions on ImageNet-1K, Places365 (Zhou et al., 2017a) and compared their linear probing performance in Table 7.

B.2. Semantics from different features

We conducted linear probing of different mean-pooled features of different types, and compared in Table 8. Sparse feature showed strongest global understanding quality.

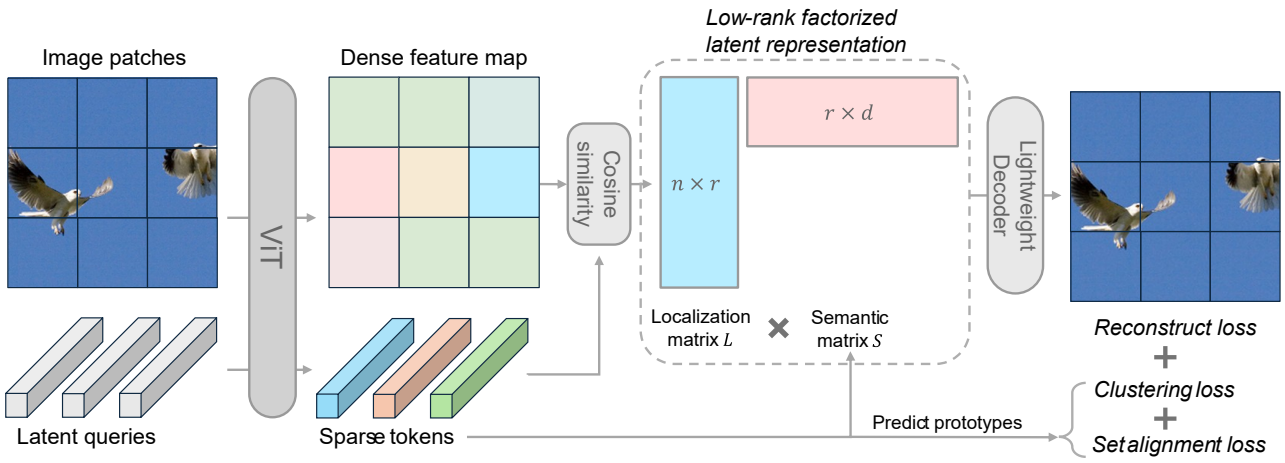

Figure 4. The STELLAR framework. We use a vanilla ViT to extract sparse tokens from an image, and model the latent representation as a low-rank matrix factorization, ensuring reconstruction of the original image. Clustering loss and set alignment loss are applied on the disentangled sparse tokens.

B.3. Concept alignment with language

Inspired by (Zhang et al., 2025), we used frozen feature from STELLAR and aligned with the text tower of CLIP (Radford et al., 2021) with a single attention pooled probing layer. The evaluation on vision language tasks with comparison to baseline models are shown in Table 9.

B.4. Finetuning

We performed finetuning for STELLAR on ImageNet-1K classification and ADE20K segmentation, and compared with baseline models. We used the same evaluation protocol as in Sec. A.2, with the backbone unfrozen and finetuned for 75 epochs. We used ViT-B for all models. The finetuning results are shown in Table 11. STELLAR showed consistent performance gain across different tasks, and close to the top model iBOT with slight difference.

B.5. Efficiency analysis

To analyze the efficiency of the STELLAR framework, we printed the processing time of the main components in the STELLAR framework with one A100 GPU at different batch sizes. Encoding the main global view of the image takes up most of the processing time, followed by encoding the masked views (8 views at 80% masking ratio) and decoding to the original image. The Sinkhorn-Knopp algorithm used for clustering and the Sinkhorn algorithm used in optimal transport matching take up much less amount of time, and their total processing time stay at similar level when increasing the batch size.

In comparison to the Sinkhorn matching algorithm we used in our experiments, we show the processing time using an alternative Hungarian matching algorithm commonly used in previous literature such as Sparse R-CNN (Sun et al., 2021), DETR (Carion et al., 2020) and MaskFormer (Cheng et al., 2021). As the implementation of the exact matching is not scalable with GPU parallelization, it's computational time increases linearly with the batch size. At batch size 64, it is already 6 times of the encoder processing, while the Sinkhorn algorithm is over 100 times faster. For this reason, we added a small entropy regularization term in the bipartite matching objective, allowing us to use the Sinkhorn algorithm for efficient matching with GPU parallelization.

C. Additional Illustration

See Fig. 4 and 5.

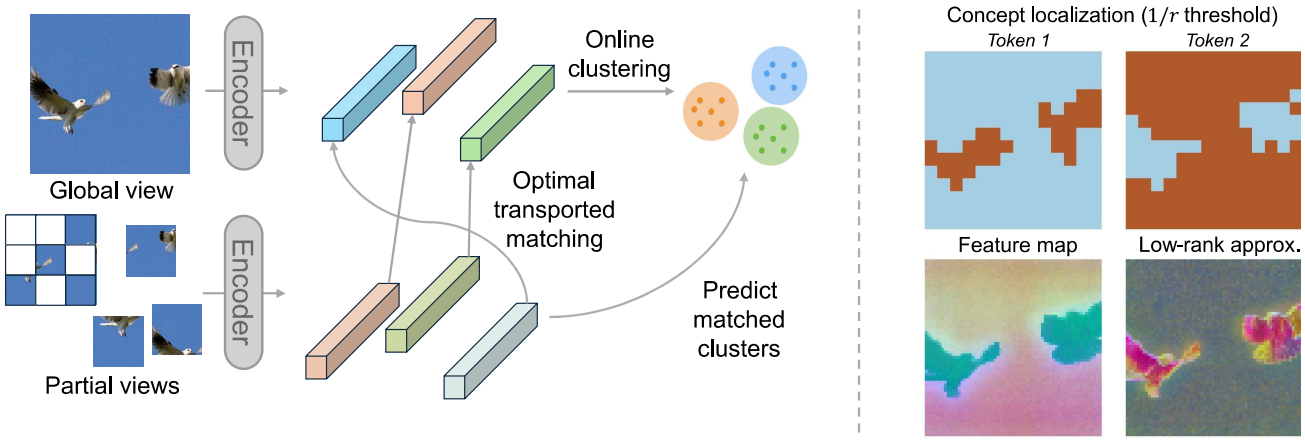

Figure 5. Left: Concept clustering and alignment workflow. Right: visualization of learned representation.

Table 6. List of baseline models and SSL method type.

Model	Reference	Method	SSL space	SSL tokens
BYOL	(Grill et al., 2020)	augmentation alignment	latent	global
MoCo v3	(Chen et al., 2021)	contrastive learning	latent	global
DINO	(Caron et al., 2021)	augmentation alignment	latent	global
MSN	(Assran et al., 2022)	masked alignment	latent	global
DenseCL	(Wang et al., 2021)	contrastive learning	latent	dense
Data2Vec	(Baevski et al., 2022)	latent MIM	latent	dense
SiameseIM	(Tao et al., 2023)	latent MIM	latent	dense
IJEPA	(Assran et al., 2023)	latent MIM	latent	dense
iBOT	(Zhou et al., 2021)	align + latent MIM	latent	global+dense
BEIT	(Bao et al., 2021)	token MIM	image	dense
SimMIM	(Xie et al., 2022)	pixel MIM	image	dense
MAE	(He et al., 2022)	pixel MIM	image	dense
SemMAE	(Li et al., 2022)	pixel MIM	image	dense
TiTok	(Yu et al., 2024)	reconstruction + clustering	image	sparse
AIM	(El-Nouby et al., 2024)	autoregressive	image	dense
DINOv2	(Oquab et al., 2023)	align + latent MIM	latent	global+dense

Table 7. Effect of pretraining data.

	linear probing acc.	
Pretraining data	ImageNet-1K	Places 365
ImageNet-1K	76.94	49.25
Places365	66.08	51.98

Table 8. Semantics in different features

Feature	sparse	cls	dense
IN-1K lin. acc (%)	73.26	72.23	72.21

Table 9. Language alignment evaluation.

	IN-1K 0-shot		MS COCO		Winoground		MMVP
	@1	@5	T2I	I2T	Text	Image	Avg.
MAE	23.18	50.43	11.28	13.46	20.75	9.00	19.26
iBOT	50.01	80.43	20.79	29.38	24.75	12.00	18.52
STELLAR	51.53	80.04	17.94	22.34	26.25	8.25	19.26
CLIP	72.7	-	43.0	59.7	30.5	11.5	20.0

Table 10. Finetuning performance in ImageNet-1K classification accuracy and ADE20K segmentation mIOU (%). We show in parentheses the gain over the respective linear probing results.

Model	ImageNet-1K Acc.	ADE20K mIOU
DINO	79.58 (+3.12)	39.22 (+12.35)
MAE	77.75 (+11.43)	40.33 (+9.42)
iBOT	80.72 (+9.14)	42.76 (+10.97)
STELLAR	80.05 (+6.78)	41.98 (+10.65)

Table 11. Processing time (s) of the main components in the STELLAR framework with one A100 GPU at different batch sizes. In comparison to the Sinkhorn matching algorithm we used in our experiments, we show the processing time using an alternative Hungarian matching algorithm commonly used in previous literature (shown in gray).

Batch size	4	8	16	32	64
Encoder	8.2×10^{-3}	9.1×10^{-3}	1.4×10^{-2}	2.0×10^{-2}	3.2×10^{-2}
Decoder	4.6×10^{-3}	6.8×10^{-3}	8.8×10^{-3}	1.2×10^{-2}	1.5×10^{-2}
Mask encoding	7.9×10^{-3}	8.9×10^{-3}	1.1×10^{-2}	1.8×10^{-2}	1.7×10^{-2}
SK clustering	3.4×10^{-4}	3.4×10^{-4}	3.4×10^{-4}	3.7×10^{-4}	3.9×10^{-4}
Sinkhorn matching	1.4×10^{-3}	1.4×10^{-3}	1.4×10^{-3}	1.4×10^{-3}	1.2×10^{-3}
Hungarian matching	5.7×10^{-3}	1.7×10^{-2}	4.0×10^{-2}	9.0×10^{-2}	1.8×10^{-1}