# Distilling Visual Information into Symbolic Representations through Self-Supervised Learning

**Victor Sebastian Martinez Pozos**
Posgrado de Ciencia e Ingeniería en Computación,
Universidad Nacional Autónoma de México
Mexico City, Mexico
martinez.victor@comunidad.unam.mx

**Ivan Vladimir Meza Ruiz**
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Mexico City, Mexico
ivanvladimir@turing.iimas.unam.mx

## Abstract

This work investigates whether self-supervised learning (SSL) can abstract visual information into structured symbolic representations without direct supervision. Drawing inspiration from information theory, we explore the use of message encoding to capture complex visual data in symbolic sequences. We aim to generate meaningful image descriptions from these representations, relying on inherent data patterns rather than explicit labels or guidance. To achieve this, we employ SSL techniques, specifically DINO, hypothesizing that the encoded symbolic sequences will closely reflect the underlying structure of the visual content. This approach aims to determine whether visual data can be effectively distilled into symbolic forms that preserve essential meaning and structure.

## 1 Introduction

In recent years, the pursuit of abstracting knowledge from high-dimensional data into discrete representations has gained significant traction [5;11;14;10], particularly in the domain of visual information. Inspired by how language encodes complex ideas symbolically, this work explores whether self-supervised learning (SSL) techniques can facilitate a similar transformation for visual data. The primary objective is to investigate how SSL can be used to convert continuous visual inputs into symbolic representations, with variable lengths—shorter ones capturing core details and longer ones providing increasingly fine-grained information, ultimately paving the way for a more structured and compositional abstraction framework.

This approach is significant because models that can autonomously abstract and structure information without direct supervision may generalize better across diverse tasks. By leveraging these structured and compositional representations, such models have the potential to advance scene understanding, reasoning, and decision-making.

To achieve this, we employ SSL, focusing on the DINO framework [1], which excels at generating coherent and consistent representations. DINO's architecture promotes alignment between a teacher and a student network, ensuring that learned representations preserve meaningful content. SSL, in this context, enables the model to uncover inherent data patterns without external guidance, making it an ideal approach for developing high-level abstractions.

Latinx in AI @ NeurIPS-24 (LXAI 2024).

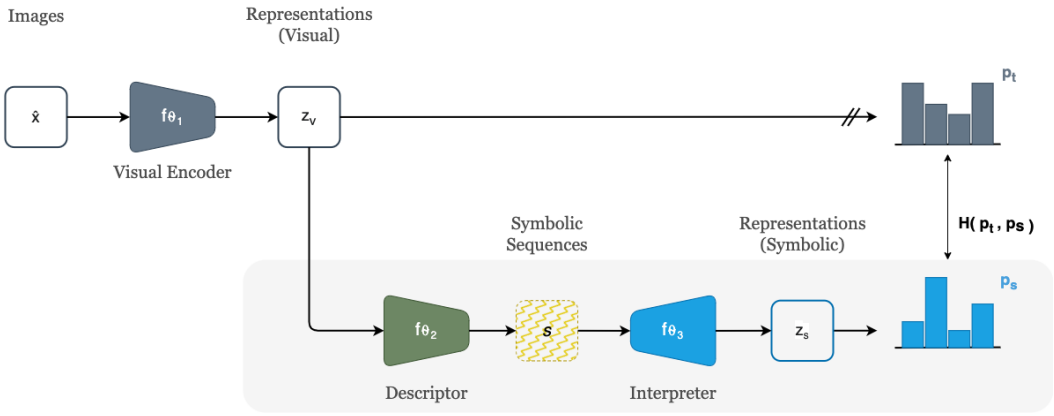


Figure 1: Schematic drawing of the teacher-student setup. The teacher model consists as usual of an encoder and projector, while the student models consist of a decoder and encoder plus the regular projector. The input images are passed to a pretrained teacher, and the representations of it are then fed to the student finally, the outputs of the projectors are compared. The student weights are then adjusted to mimic the output of the teacher. In our experiments, we work under the assumption of an existing visual encoder and focus solely on training the projector layer of the teacher using EMA while keeping the rest of it frozen.

Our preliminary experiments aim to determine whether SSL can generate symbolic sequences that effectively reflect the underlying structure of visual data. Although challenges remain—particularly in balancing compression with fidelity and ensuring interpretability across diverse contexts—this work lays the groundwork for future efforts toward scalable, high-level visual understanding, offering insights into how structured symbolic representations can be learned without explicit supervision.

## 2 SSL with Symbolic Representations

In self-supervised learning (SSL), knowledge distillation [7] is employed to transfer knowledge from a teacher model $\mathbf{T}$ to a student model $\mathbf{S}$. The goal is for the student to replicate the teacher's understanding of the data by minimizing the divergence between their feature representations.

The teacher-student loss function is defined as:

$$\mathcal{L}_{\text{KD}}(\mathbf{S}, \mathbf{T}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[d\left(\mathbf{S}(\mathbf{x}), \mathbf{T}(\mathbf{x})\right)\right]$$

where $\mathbf{S}(\mathbf{x})$ and $\mathbf{T}(\mathbf{x})$ represent the feature representations or predictions of the student and teacher models, respectively. The function $d(\cdot, \cdot)$ is a distance metric (commonly the Kullback-Leibler divergence or mean squared error) that quantifies the difference between the student's and teacher's outputs.

In our framework, we adopt a teacher-student architecture based on the DINO paradigm, as shown in Fig. 1. Here, the teacher model $\mathbf{T}$ generates stable target representations for the student $\mathbf{S}$, and the student is trained to align its output with the teacher's latent representations.

**Teacher Network**: The teacher network follows a standard setup with an encoder module $E_T$ and a projector head $P_T$. It receives an input image $x$, which is processed by the encoder to generate visual representations $z_t$. These representations are then projected into a joint distribution space:

$$\mathbf{p}_t = P_T(z_t),$$

Since the encoder is assumed to be pre-trained, we update only the projector head in the teacher network using an exponential moving average (EMA) of the student projector weights:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$$

**Student Network**: The student network introduces a descriptor-interpreter dynamic and consists of a decoder module (descriptor) $D_S$, an encoder module (interpreter) $E_S$, and a projector head $P_S$. The student model operates in two phases: symbolic sequence generation and embedding alignment.

- **Symbolic Sequence Generation**: To generate descriptions of varying levels of detail, the descriptor $D_S$ autoregressively transforms the teacher's visual representations $z_t$ into symbolic sequences $s_s$ using cross-attention mechanisms. These sequences represent high-level semantic abstractions, capturing key features of the input. By generating symbolic sequences of different lengths for the same scene, shorter sequences focus on broad semantic features, while longer sequences capture more specific details. This approach encourages the student model to learn a general-to-specific behavior, enabling it to adjust to different levels of abstraction in the data.

$$\mathbf{s_s} = D_S(\mathbf{z}_t),$$

- **Discretization and Re-Embedding**: The symbolic sequence $\mathbf{s}_s$ is discretized into symbolic tokens. These tokens, representing key aspects of the image, are encoded back into continuous embeddings through the interpreter, which generates symbolic representations $z_s$. Finally these representations are passed through the student's projector into a joint distribution space:

$$\mathbf{p}_s = P_S(E_S(\mathbf{s}_s)),$$

**Loss Calculation**: The loss function is designed to guide the student model in learning from both continuous and symbolic representations. Building on the DINO loss, we introduce a granularity term that accounts for varying levels of detail in the symbolic representations. This term adapts the local-to-global strategy from the DINO framework, encouraging the student to align its representations with those of the teacher across different levels of abstraction. The granularity factor allows the student to focus on both high-level and more detailed features of the input.

The overall loss function $\mathcal{L}_{SSL}$ is defined as:

$$\mathcal{L}_{SSL} = \sum_{i=1}^{V} \sum_{j=1}^{D} \lambda^j \times \mathcal{H}(\mathbf{p}_t^{(i)}, \mathbf{p}_s^{(i,j)})$$

where $\mathcal{H}(\mathbf{p}_t^{(i)}, \mathbf{p}_s^{(i,j)})$ is the cross-entropy between the teacher's visual representations $\mathbf{p}_t$ and the student's symbolic representations at different granularities $\mathbf{p}_s$, with $\lambda^j$ acting as a scaling factor for each level of detail.

**Implementation details**. We pretrain the models on the CIFAR-10 dataset[9] without labels. The models are trained using parameters mostly in line with those reported in[1], with a few exceptions: the batch size was set to 64, a single GPU was used, the projection size was 8192, and the learning rate was adjusted. For the teacher encoder, we use a pretrained ViT-B/16[4], while for the student descriptor we use a custom decoder with 8 heads and 6 layers and for the student interpreter a custom encoder with 6 heads and 12 layers. Additionally, the method uses only four global views as inputs and generates sequences of up to eight symbols.

## 3 Symbolic sequences generation

**Role of the Quantization Process**: In evaluating the performance of our method, we explored three distinct approaches for generating symbolic sequences: 1) vector quantization (VQ)[11], 2) Gumbel softmax (GS)[8], and 3) sharpened softmax (SS)[13]. As shown in Fig. 2, all three techniques demonstrated improvements in abstracting information from the data. However, their performance eventually plateaued, indicating diminishing returns beyond a certain point. Among these methods, Gumbel softmax yielded the highest evaluation scores, suggesting a stronger ability to capture diverse representations. In contrast, the sharpened softmax models exhibited significant instability, introducing excessive noise, with some variations collapsing during training. A closer analysis of the performance gap between GS and VQ suggests that the entropy regularization in GS might encourage

greater diversity in the generated sequences, whereas VQ, by design, produces sequences in a more deterministic manner, without such regulatory mechanisms.

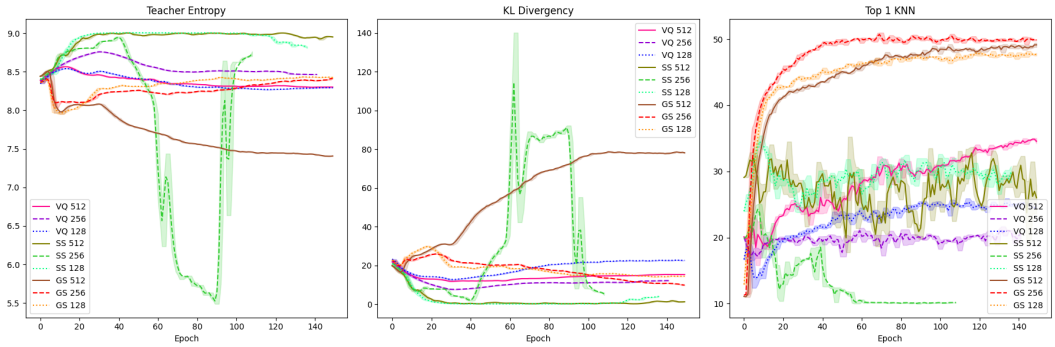


Figure 2: Training process of nine variations of our method, including three discretization variations with varied vocabulary sizes in symbolic descriptions. From left to right: (a) shows the teacher entropy over training steps; (b) displays the KL divergence between teacher and student distributions; (c) presents the evaluation performance using a k-NN metric across the different variations.

**Evaluation of symbolic sequences.** To evaluate learned representations, we employ probing techniques such as linear classification and k-nearest neighbors (k-NN), following standard protocols[1;3;2]. In linear evaluation, a classifier is trained on frozen features extracted from the student, providing insight into representational quality without additional learning[6]. For k-NN, we freeze the model and identify each test image's nearest neighbors in the training set, assessing the learned features' effectiveness[12]. Both probing methods are applied to the CIFAR-10 validation set, reporting top-1 and top-5 accuracy for a comprehensive evaluation. For k-NN in particular, the reported metrics are averaged over four different choices of $n$ (number of neighbors), ensuring a robust and balanced evaluation of the learned features.

Table 1: KNN and Linear Probing Results

| **Model** | **KNN** | | **Linear** | |
|---|---|---|---|---|
| | **Top1** | **Top5** | **Top1** | **Top5** |
| VQ 512 | 34.2825 | 88.365 | 0.3365 | 0.9030 |
| VQ 256 | 20.7375 | 78.810 | 0.2082 | 0.7992 |
| VQ 128 | 25.3150 | 80.0675 | 0.2202 | 0.8299 |
| SS 512 | 31.6525 | 88.640 | 0.3335 | 0.9108 |
| SS 256 | 10.1450 | 50.2425 | 0.0995 | 0.4976 |
| SS 128 | 30.7925 | 86.6575 | 0.2868 | 0.8982 |
| GS 512 | 49.1250 | 93.810 | 0.5248 | **0.9686** |
| GS 256 | **50.0150** | **94.185** | **0.5301** | 0.9674 |
| GS 128 | 47.3075 | 91.105 | 0.4983 | 0.9456 |

**Role of Vocabulary Size**. We investigated the influence of vocabulary size on training and performance using symbol sets of 128, 256, and 512 elements. The results, summarized in Fig. 2, revealed that while certain metrics improved with larger vocabularies, the trend was inconsistent, and overall performance across models of the same family was similar. This raises the possibility that the observed effects stem from the interaction between vocabulary size and sequence length, as the effective search space grows exponentially with these parameters. Further experiments with reduced vocabulary sizes (e.g., 64, 32, or 16 symbols) could shed light on this relationship.

## 4 Discussion

This study demonstrates the potential of self-supervised learning (SSL) in generating structured symbolic representations from visual data. By exploring various generation strategies, including

Gumbel softmax, vector quantization, and sharpened softmax, we found that Gumbel softmax promotes greater diversity and stability in the generated sequences. However, the effect of increasing vocabulary size on representation quality was inconsistent, with larger vocabularies sometimes introducing redundancy rather than enhancing performance. While all methods showed promise in abstracting visual information, challenges remain in balancing sequence detail and efficiency, highlighting the need for further refinement in these approaches.

Future research will expand on these findings by testing these methods across a broader range of datasets to assess their generalizability. Ablation studies will also be conducted to explore the impact of different variations in the student components on performance. Additionally, we plan to investigate how these symbolic sequences can be applied to downstream tasks such as visual reasoning, where structured representations learned through SSL could offer significant benefits. This work contributes to the exploration of symbolic abstraction in visual data, which may pave the way for more advanced and interpretable AI models in the future.

## Acknowledgments

The authors thank the "Proyectos de investigación en la Nube UNAM-AWS" program, which supports the research by providing computing time.

Additionally, the authors would like to thank the Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), Mexico, for supporting this research through the "Beca Nacional para Estudios de Posgrado".

## References


[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

[3] X. Chen and K. He. Exploring simple siamese representation learning, 2020. URL https://arxiv.org/abs/2011.10566.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[5] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[6] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

[7] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

[8] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

[9] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research), 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html. Dataset.

[10] N. Mu, A. Kirillov, D. Wagner, and S. Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer Nature Switzerland, October 2022.

[11] A. Van Den Oord and O. Vinyals. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[12] Z. Wu, Y. Xiong, S. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. URL https://arxiv.org/abs/1805.01978.

[13] X. Zhang, F. X. Yu, S. Karaman, W. Zhang, and S.-F. Chang. Heated-up softmax embedding, 2018. URL https://arxiv.org/abs/1809.04157.

[14] Y. Zhong, Z.-Y. Hu, M. R. Lyu, and L. Wang. Beyond embeddings: The promise of visual table in visual reasoning, 2024. URL https://arxiv.org/abs/2403.18252.