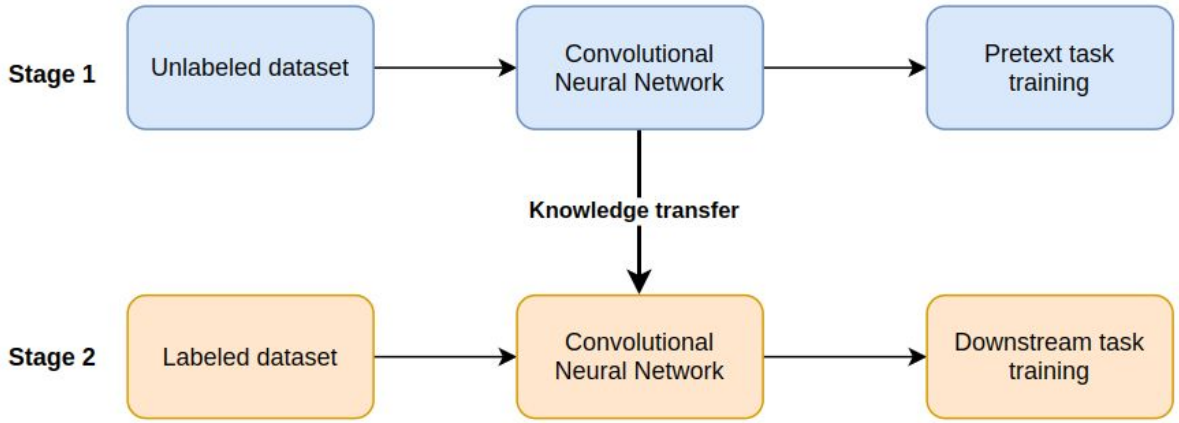
# Self-supervised learning
# SSL

***semiDPS 2025***

*Sesión 2*

# Retomemos...




***Esquema SSL***

SSL se enfoca en aprender **características discriminativas** sobre grandes cantidades de datos no etiquetados de la siguiente manera:

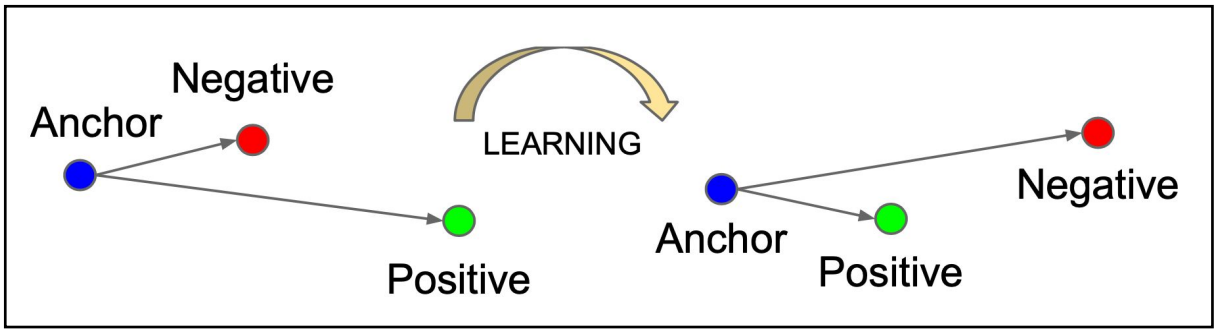
- ***Pretext task***: generar pseudo-etiquetas según el tipo de dato, *SIN NECESIDAD DE ETIQUETAS*. Se termina con un modelo pre-entrenado para un problema diferente al objetivo end-to-end.

- ***Downstream task***: transferir el modelo aprendido en la etapa anterior para la tarea específica *CON POCOS DATOS ETIQUETADOS.*

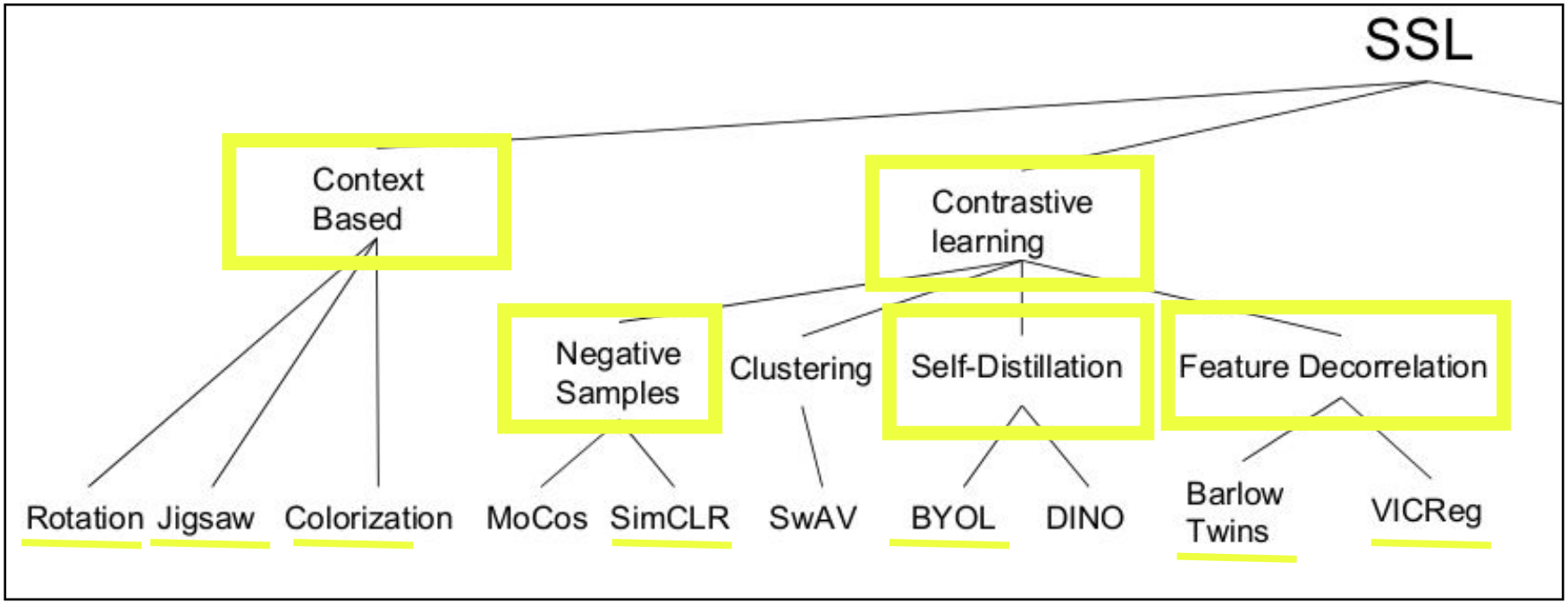
# Retomemos...

Context-based pretext task


Contrast Learning pretext task

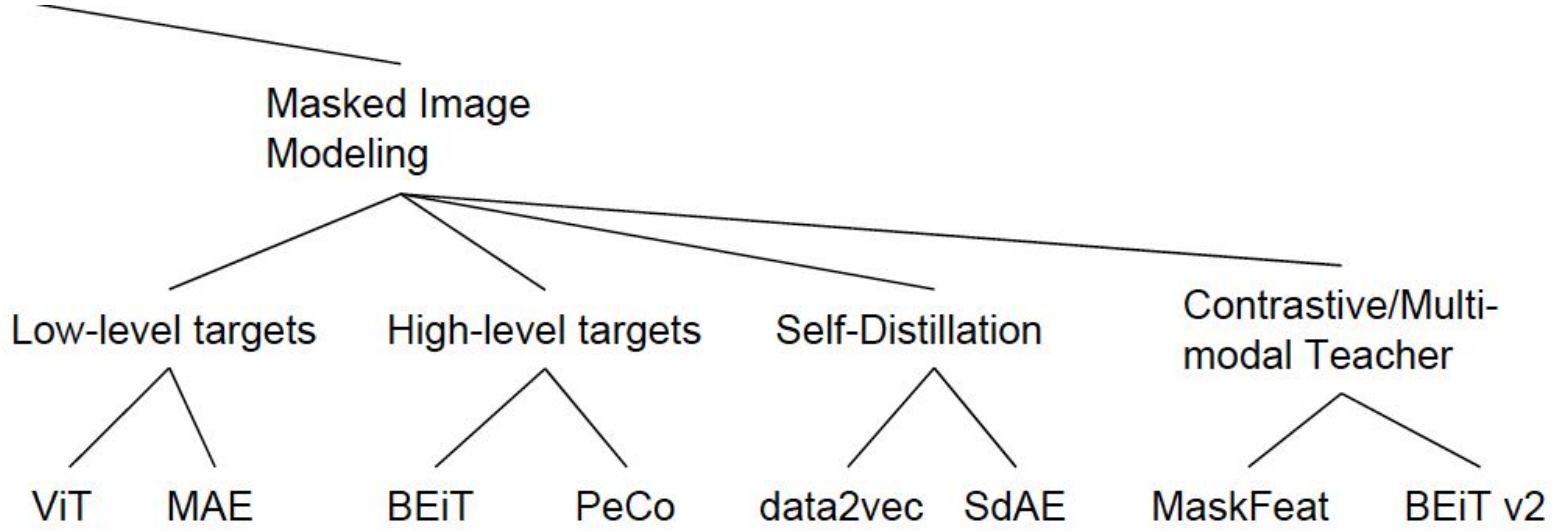


# Vimos Context Based  y Contrastive Learning...



# Masked Image Modeling

**SSL**



# Masked Image Modeling

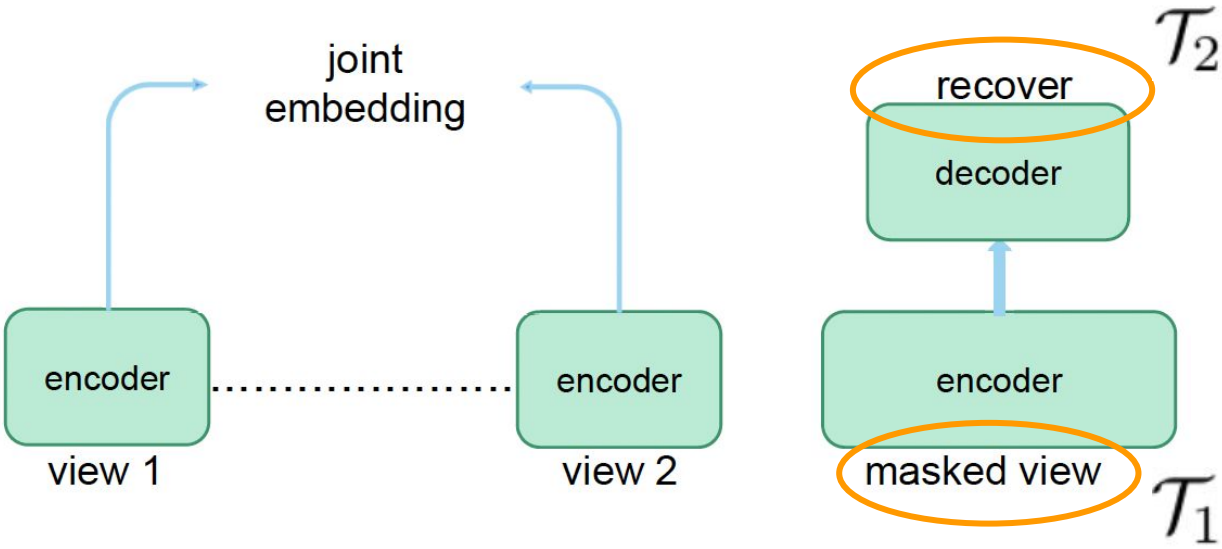
$$\text{MIM} := \mathcal{L}\left(\mathcal{D}\left(\mathcal{E}\left(\mathcal{T}_1\left(I\right)\right)\right), \mathcal{T}_2\left(I\right)\right)$$

$\mathcal{E}$ denotes the encoder, $\mathcal{D}$ denotes the decoder,

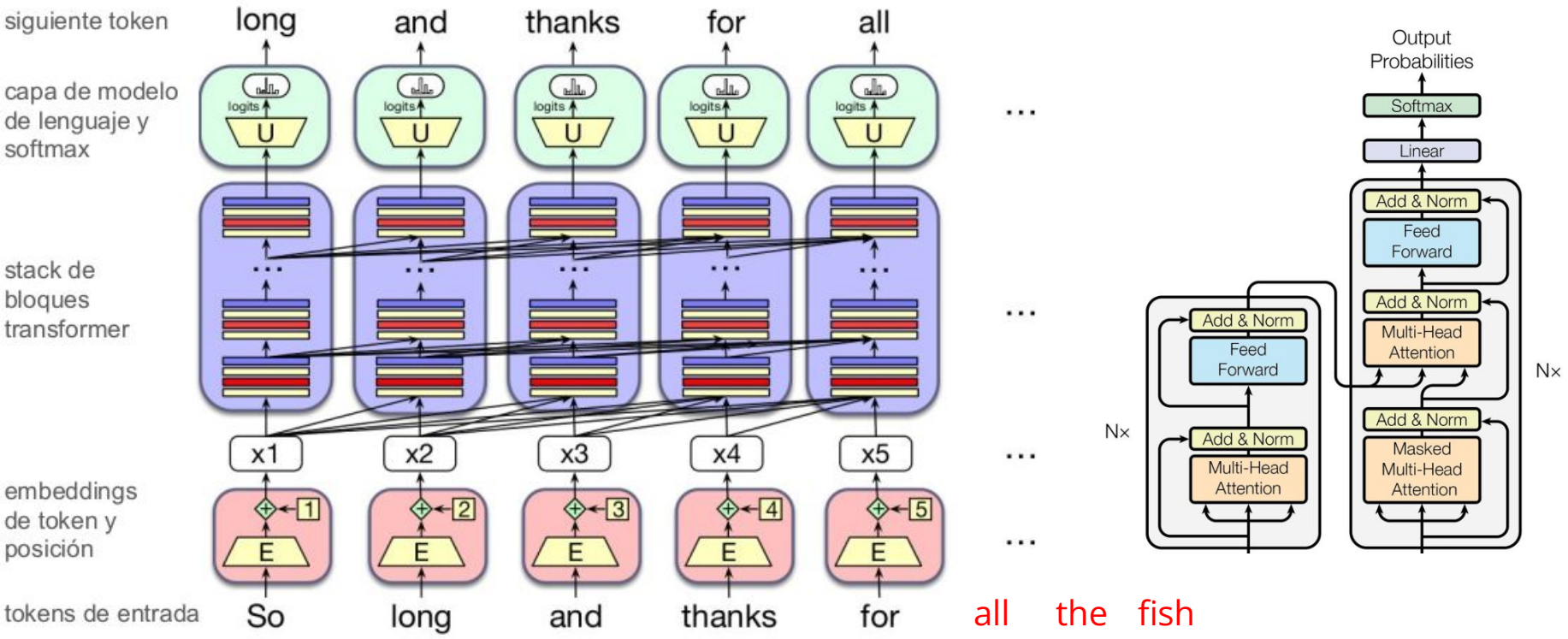
$\mathcal{T}_{1,2}$ transformaciones

1. aplicada antes de la imagen al encoder (**suele ser enmascaramiento**)
2. transformación objetivo de todo el sistema encoder-decoder

# CL vs MIM


joint
embedding
encoder
encoder
view 1
view 2
$\mathcal{T}_2$
recover
decoder
encoder
masked view
$\mathcal{T}_1$

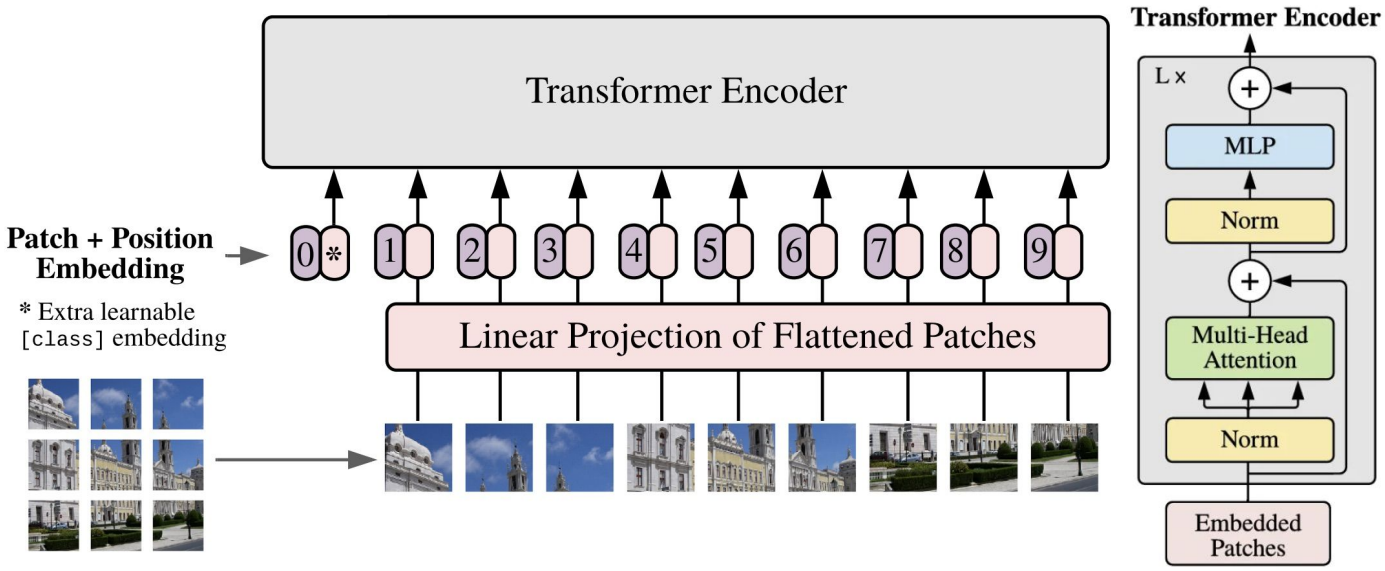
# MIM - Preámbulo: Transformer



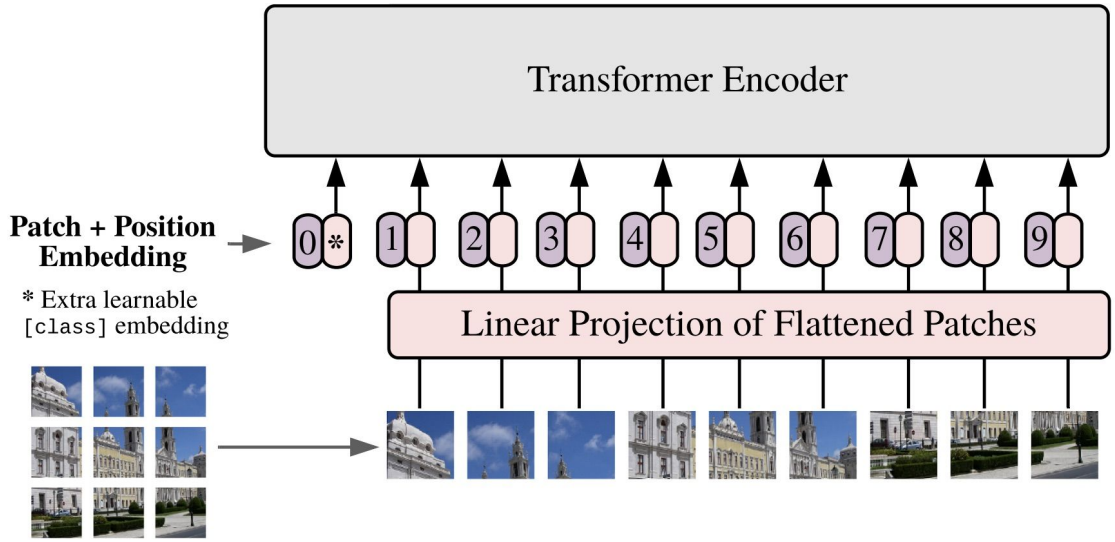
# MIM - Low Level Targets - ViT

**Patchify the Image:**
Split the image into smaller patches. ViT splits this image into smaller patches, say 16x16 patches.

**Linear Projection:**
Flatten each patch and project it into a vector (token). These patches are flattened into a vector, like turning a mini image into a word in a sentence.



# MIM - Low Level Targets - ViT




$$\text{MIM} := \mathcal{L}\left(\mathcal{D}\left(\mathcal{E}\left(\mathcal{T}_1\left(I\right)\right)\right), \mathcal{T}_2\left(I\right)\right)$$

$\mathcal{T}_{1,2}$ transformaciones

1. enmascarar “futuros” tokens
2. operaciones ***low-level*** en *I*

**a) predicting only the mean**, 3bit color (i.e., 1 prediction of 512 colors)
**b) predicting a 4 × 4 downsized version** of the 16 × 16 patch with 3bit colors in parallel
**c) regression on the full patch** using L2 (i.e., 256 regressions on the 3 RGB channels)

# MIM - Low Level Targets - ViT-MAE

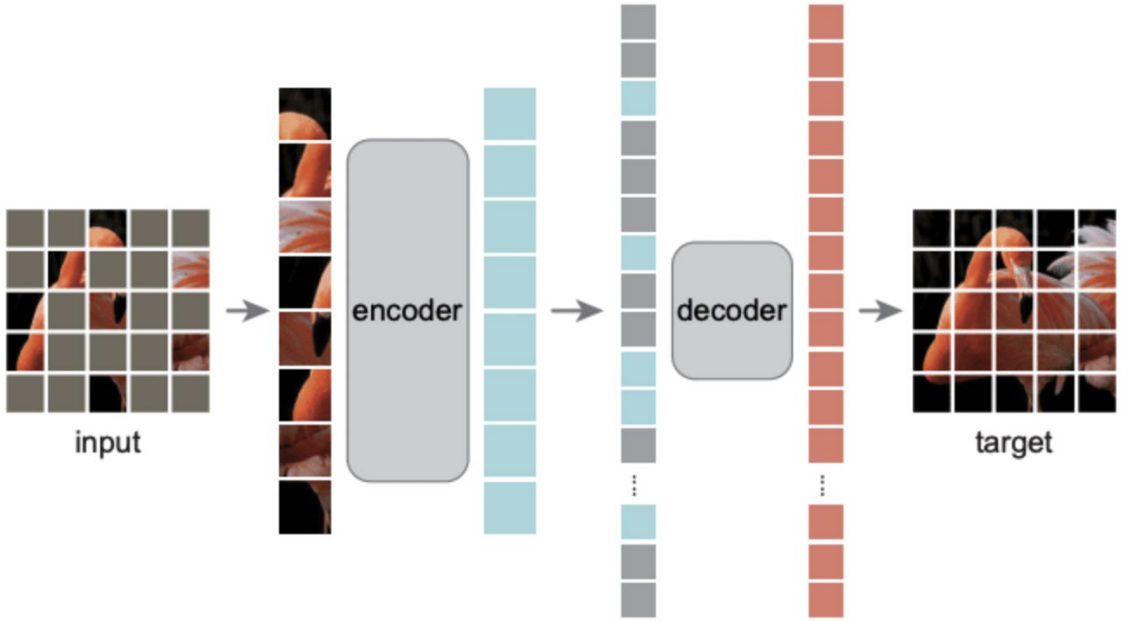
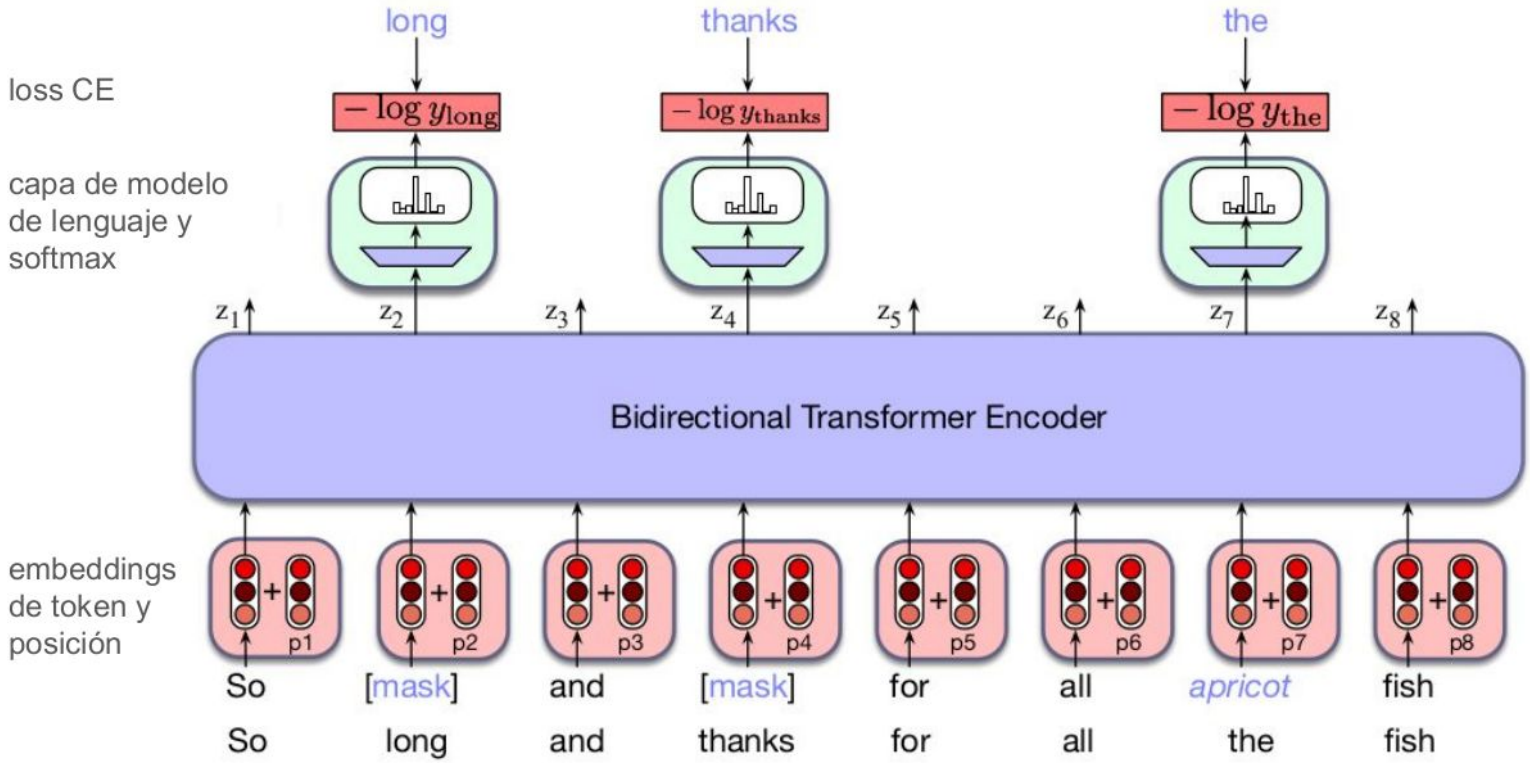



Figura 6: Arquitectura de ViT-MAE

# MIM - Preámbulo: BERT



# MIM - High Level Targets - BEiT

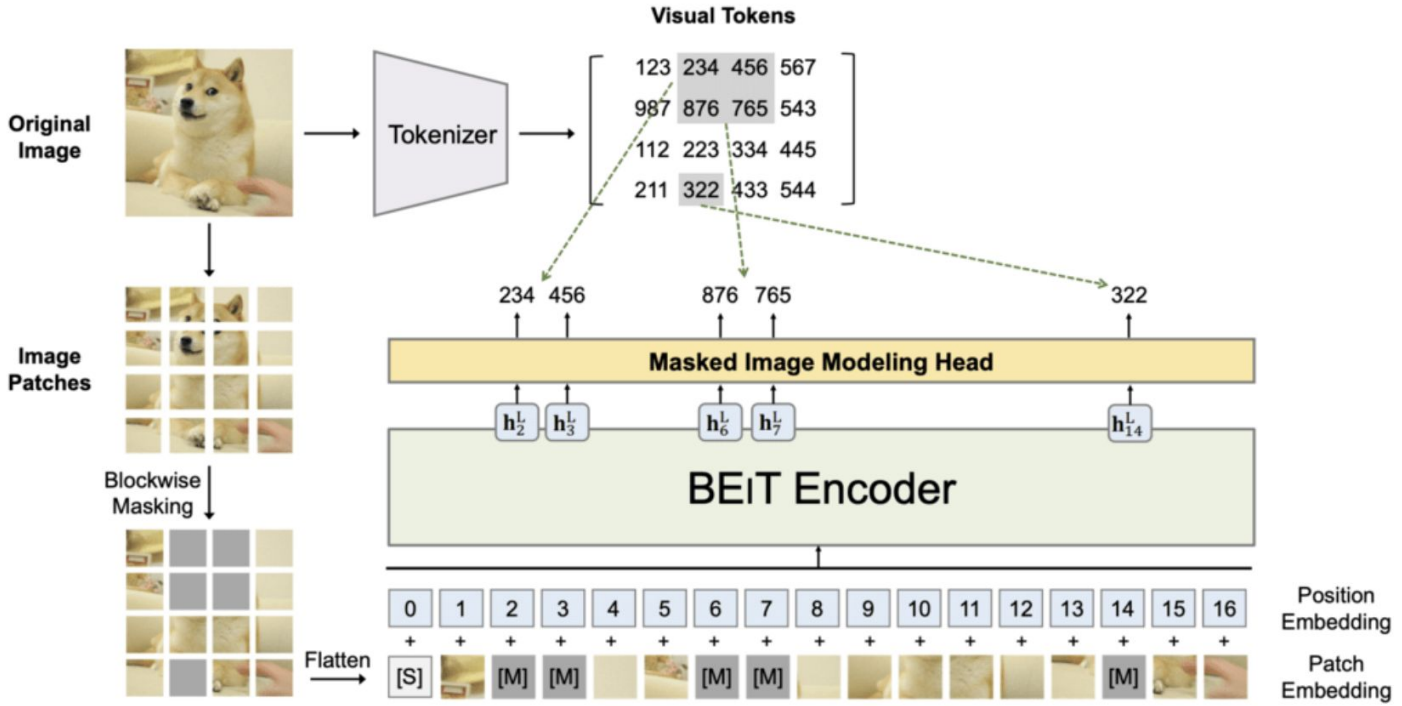



Figura 5: Arquitectura BEiT

# MIM - High Level Targets - BEiT

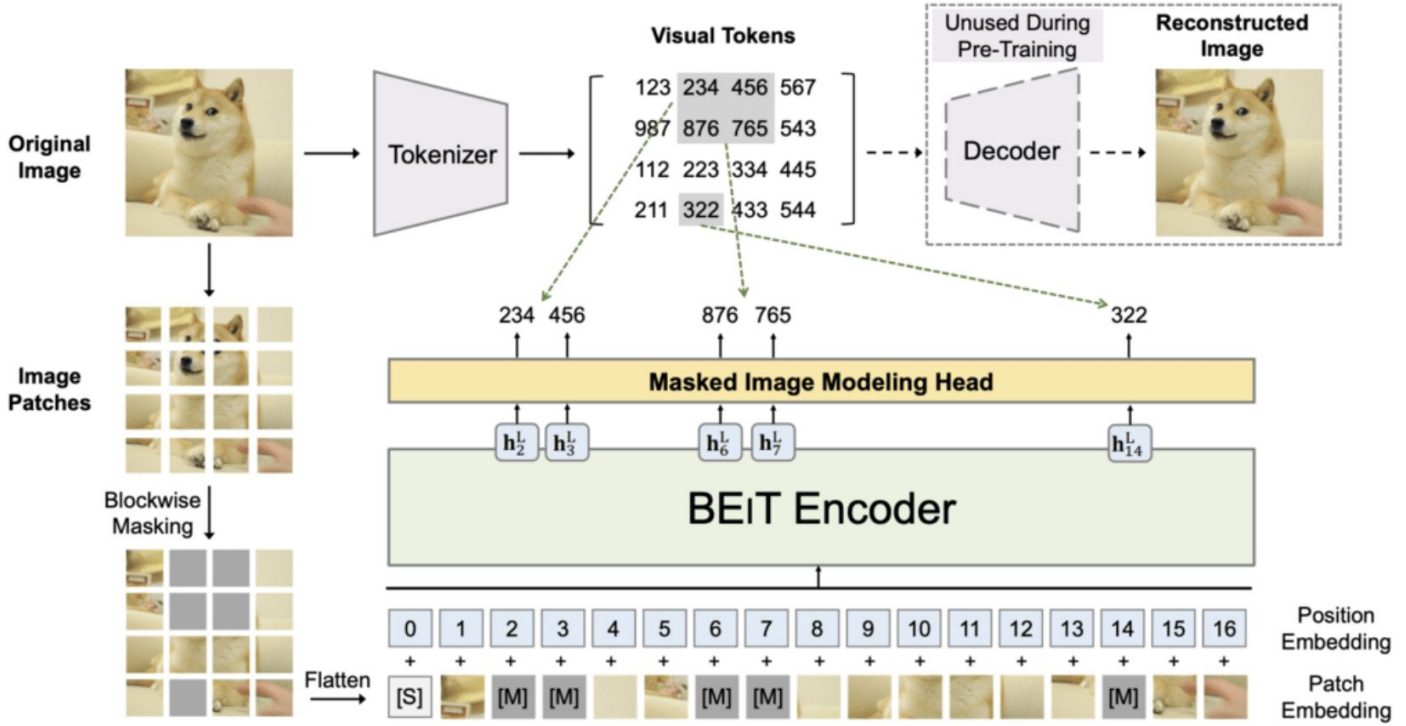
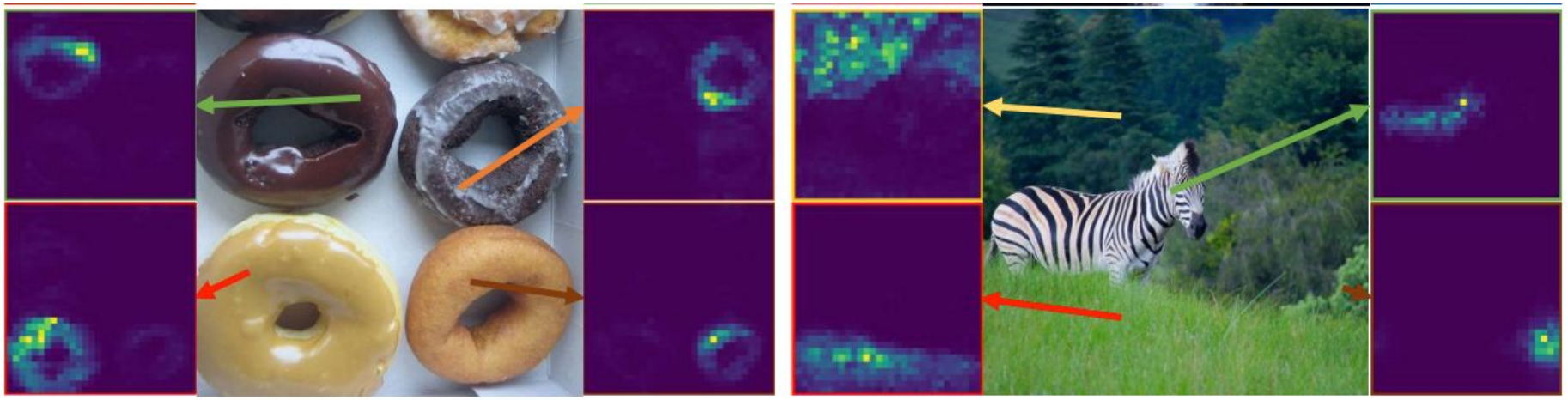



Figura 5: Arquitectura BEiT

# MIM - High Level Targets - BEiT

*"... BEiT learns to distinguish semantic regions and object boundaries, although without using any human annotation"*



**Attention scores computed via query-key product in the last layer**. For each reference point, we use the corresponding patch as query, and show which patch it attends to.

# MIM - High Level Targets - BEiT

$\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ into

$\boldsymbol{z} = [z_1, \ldots, z_N] \in \mathcal{V}^{h \times w}$

vocabulary $\mathcal{V} = \{1, \ldots, |\mathcal{V}|\}$

contains discrete token indices.

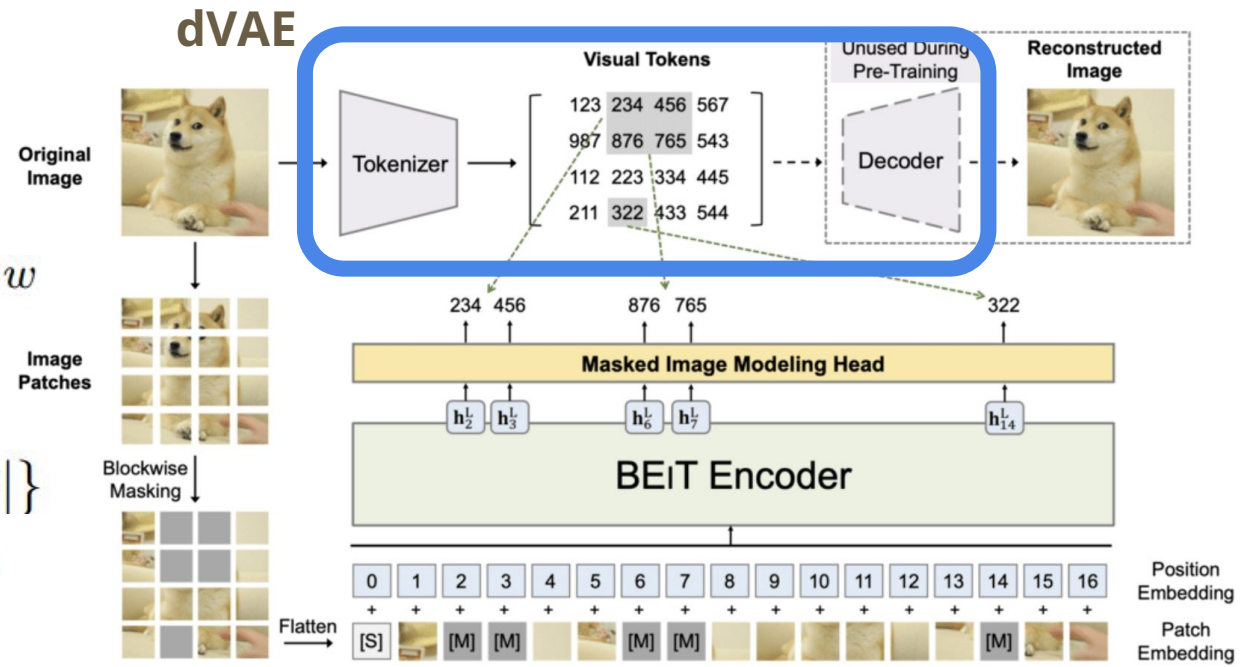



Figura 5: Arquitectura BEiT

# MIM - High Level Targets - BEiT

$x_i, \tilde{x}_i$ original y enmascarada respect.

**(ELBO)**

$$\sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \log p(x_i|\tilde{x}_i) \geq \sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \Big( \underbrace{\mathbb{E}_{z_i\sim q_\phi(\mathbf{z}|x_i)}[\log p_\psi(x_i|z_i)]}_{\text{Visual Token Reconstruction}} - D_{\mathrm{KL}}[q_\phi(\mathbf{z}|x_i), p_\theta(\mathbf{z}|\tilde{x}_i)] \Big)$$

*tokenizer* *decoder* *transformer*

$$\geq \sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \Big( \underbrace{\mathbb{E}_{z_i\sim q_\phi(z|x_i)}[\log p_\psi(x_i|z_i)]}_{\text{Stage 1: Visual Token Reconstruction}} + \underbrace{\log p_\theta(\hat{z}_i|\tilde{x}_i)}_{\text{Stage 2: Masked Image Modeling}} \Big)$$

$$\hat{z}_i = \arg\max_z q_\phi(z|x_i)$$

**(dVAE)**

# Masked Image Modeling

SSL