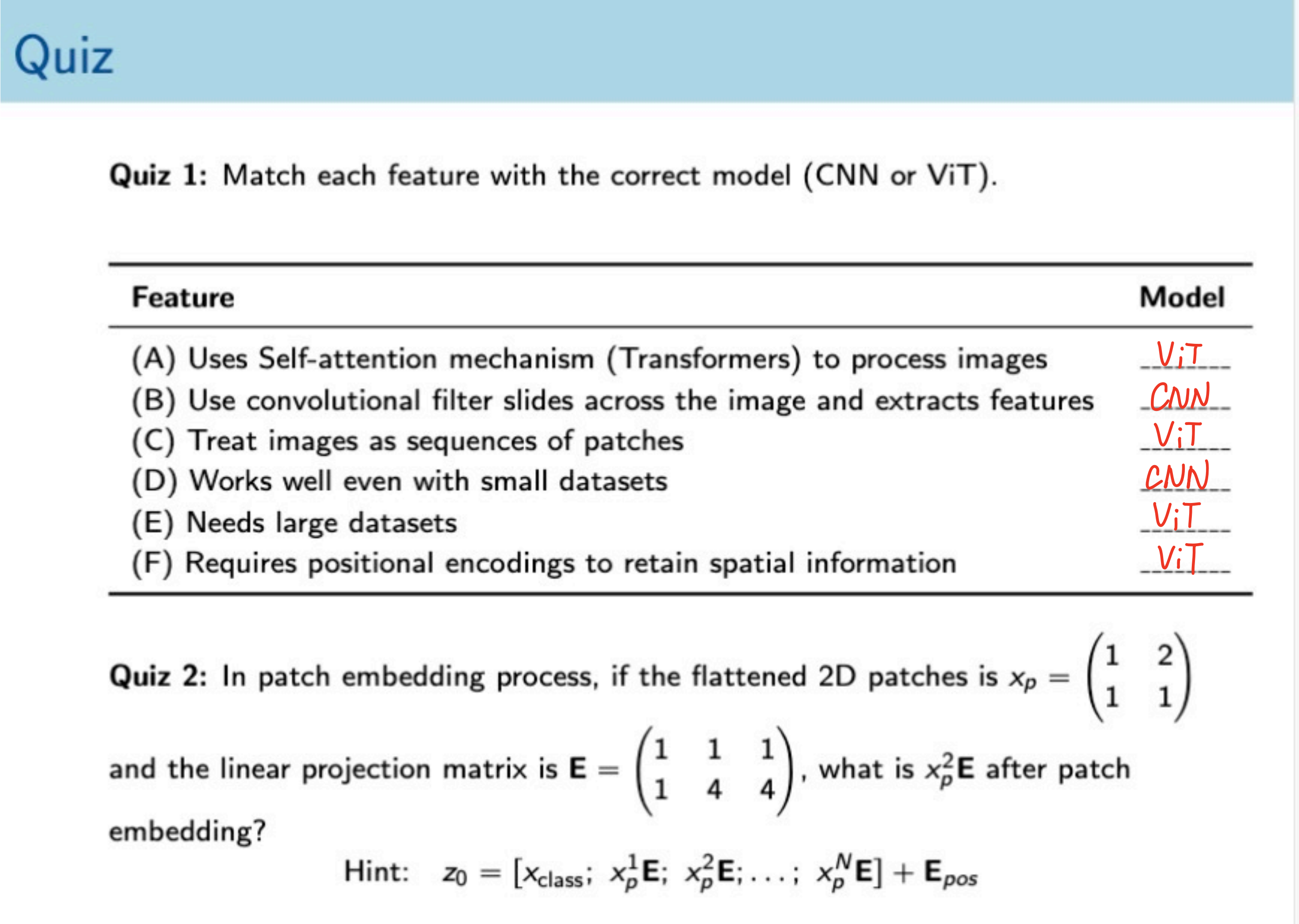

Quiz

Quiz 1: Match each feature with the correct model (CNN or ViT).

Feature	Model
(A) Uses Self-attention mechanism (Transformers) to process images	ViT
(B) Use convolutional filter slides across the image and extracts features	CNN
(C) Treat images as sequences of patches	ViT
(D) Works well even with small datasets	CNN
(E) Needs large datasets	ViT
(F) Requires positional encodings to retain spatial information	ViT

Quiz 2: In patch embedding process, if the flattened 2D patches is $x_p = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$ and the linear projection matrix is $\mathbf{E} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 4 \end{pmatrix}$, what is $x_p^2\mathbf{E}$ after patch embedding?

$$\text{Hint:} \quad z_0 = [x_{\text{class}};\ x_p^1\mathbf{E};\ x_p^2\mathbf{E}; \ldots;\ x_p^N\mathbf{E}] + \mathbf{E}_{pos}$$

$$x_p^2 E = (1 \quad 1)\begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 4 \end{pmatrix} = (2 \quad 5 \quad 5)$$