

Origin Visual Token Encoding in UMMs

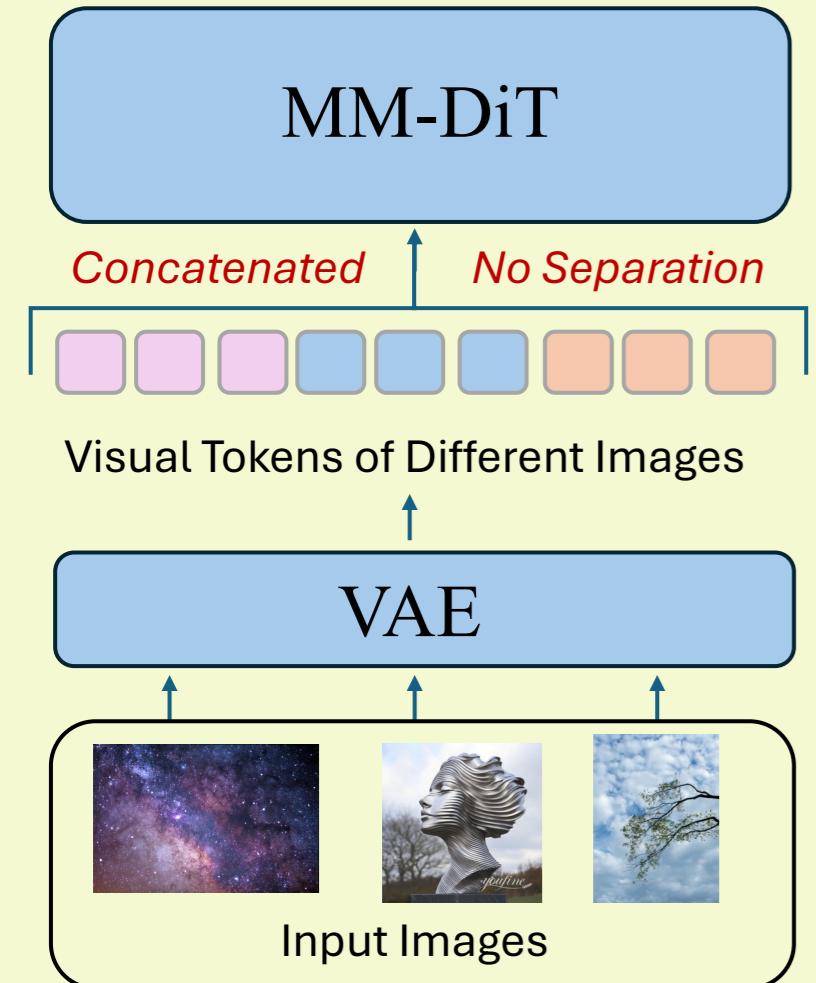
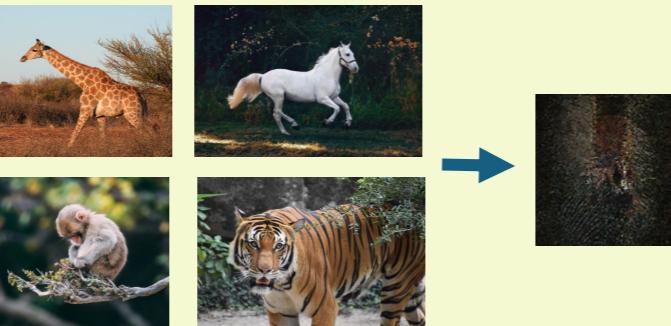


Image Identity Confusion



Replace garden hose in image 1 with a vibrant, delicate floral headpiece in image 2.

Weak Generalization to Variable Inputs



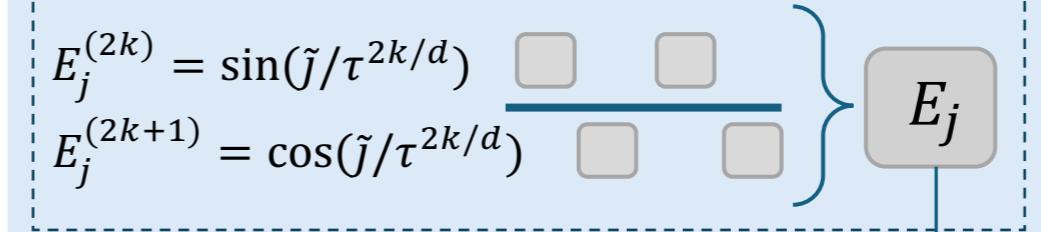
Put the horse in image 2 and tiger in image 4 to the scene in the image 1, add the monkey in image 3 next to the tiger.

Generalized Absolute Image Indexing

Sinusoidal Image Index Embedding

$$E_j^{(2k)} = \sin(\tilde{j}/\tau^{2k/d})$$

$$E_j^{(2k+1)} = \cos(\tilde{j}/\tau^{2k/d})$$



$$E_1$$

$$+$$



$$v_1^1$$

$$E_2$$

$$+$$

$$v_1^2$$

$$E_3$$

$$+$$

$$v_1^3$$

$$+$$

$$v_2^1$$

$$+$$

$$v_2^2$$

$$+$$

$$v_2^3$$

$$<\text{sep}>$$

Learnable Visual Separator Token

MM-DiT

