A Text Operation View

Text Representation Generator

Text Representation Generator

Truncate if prompt consists of more than 77 tokens
Pad with <end> if less than 77 tokens

Tokenizer → 77 tokens → **Text Encoder** → 768D vector for each token

<start> a cute and adorable bunny,

<start>	-0.39	0.02	-0.05	...
a	0.03	-1.33	0.31	...
cute	-0.55	0.11	-0.01	...

B Text-image Linkage Explanation

Image Encoder   Text Encoder   Hybrid Text-Image Space

0.5	1.3	-0.1	2.8

photo of white dog

0.4	1.1	0.4	3.0

-4.1	0.0	1.5	5.4

drawing of yellow flower

-3.8	0.2	1.1	5.2

Stable Diffusion uses CLIP's **text encoder** to connect texts with images. For an image-caption pair, CLIP's text encoder converts the caption's tokens into a vector that is close to the image's vector (produced by CLIP's image encoder) in hybrid text-image space.