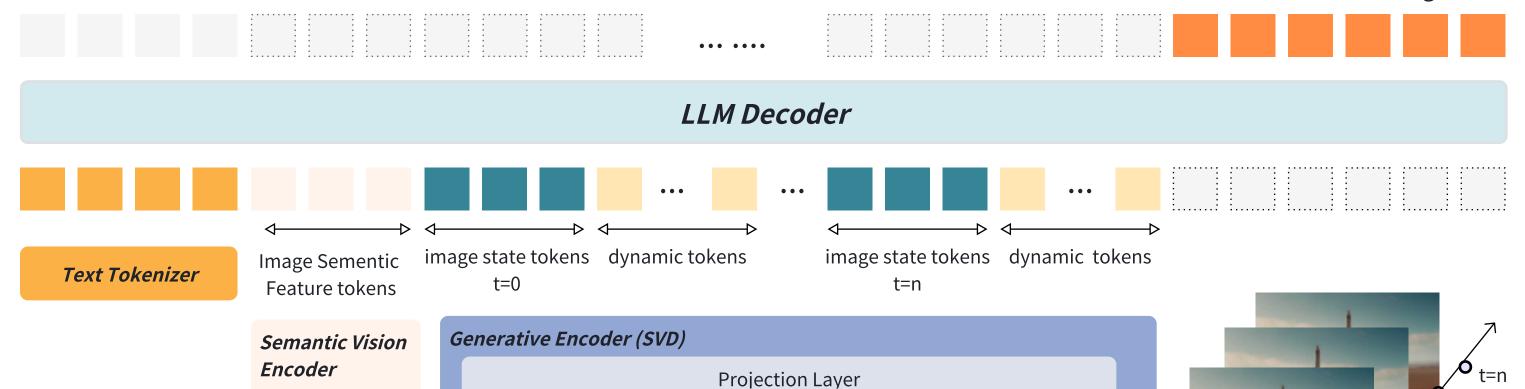The rocket launching...

LLM Decoder

Text Tokenizer

Image Sementic Feature tokens

image state tokens $t=0$

dynamic tokens

image state tokens $t=n$

dynamic tokens

Semantic Vision Encoder

SigLip

Generative Encoder (SVD)

Projection Layer

$t=n$

$t=2$

$t=1$