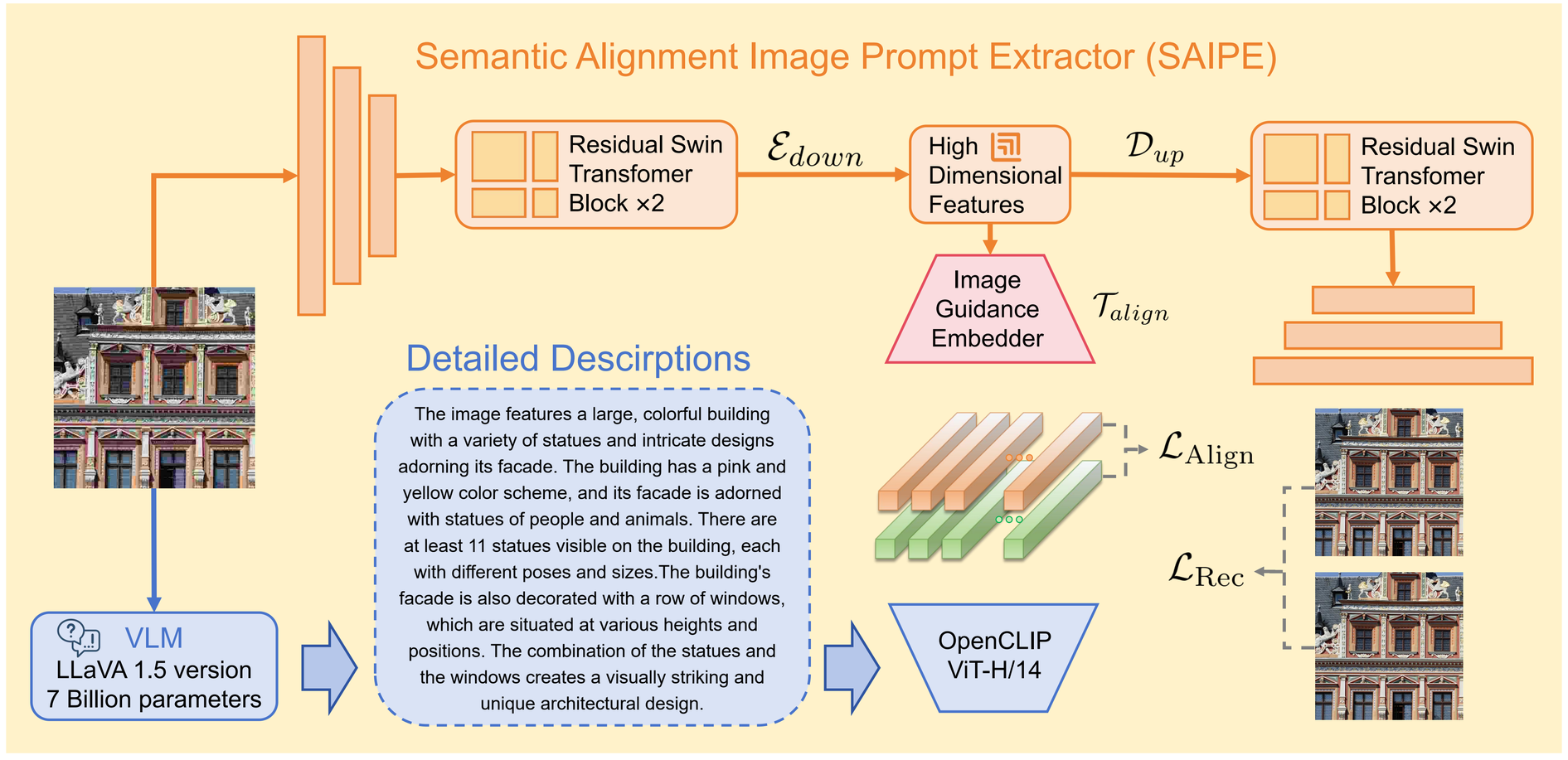
Semantic Alignment Image Prompt Extractor (SAIPE)
Residual Swin Transfomer Block ×2
$\mathcal{E}_{down}$
High Dimensional Features
$\mathcal{D}_{up}$
Residual Swin Transfomer Block ×2
Image Guidance Embedder
$\mathcal{T}_{align}$
Detailed Descirptions
The image features a large, colorful building with a variety of statues and intricate designs adorning its facade. The building has a pink and yellow color scheme, and its facade is adorned with statues of people and animals. There are at least 11 statues visible on the building, each with different poses and sizes.The building's facade is also decorated with a row of windows, which are situated at various heights and positions. The combination of the statues and the windows creates a visually striking and unique architectural design.
VLM
LLaVA 1.5 version
7 Billion parameters
OpenCLIP ViT-H/14
$\mathcal{L}_{\text{Align}}$
$\mathcal{L}_{\text{Rec}}$

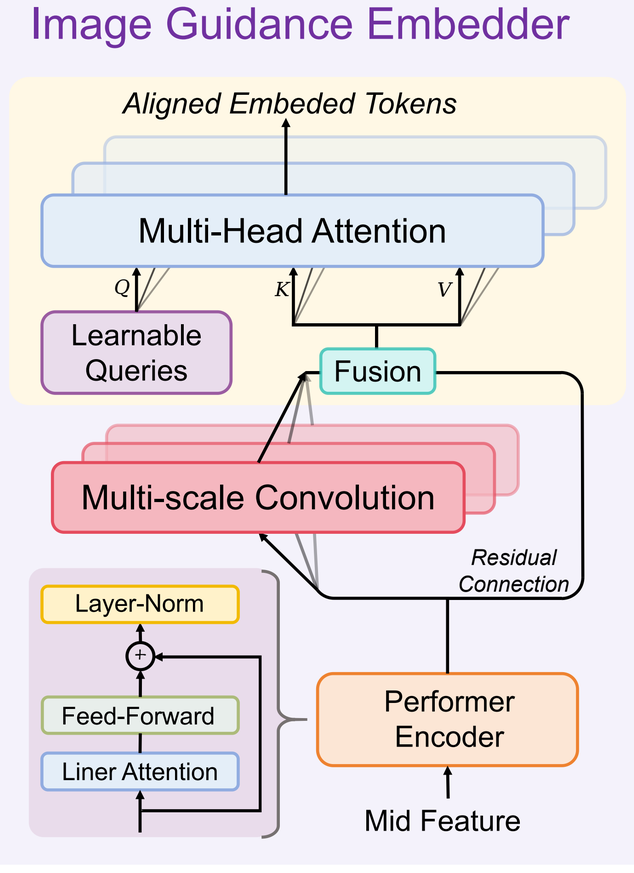
Image Guidance Embedder
Aligned Embeded Tokens
Multi-Head Attention
Q
K
V
Learnable Queries
Fusion
Multi-scale Convolution
Residual Connection
Layer-Norm
Feed-Forward
Liner Attention
Performer Encoder
Mid Feature