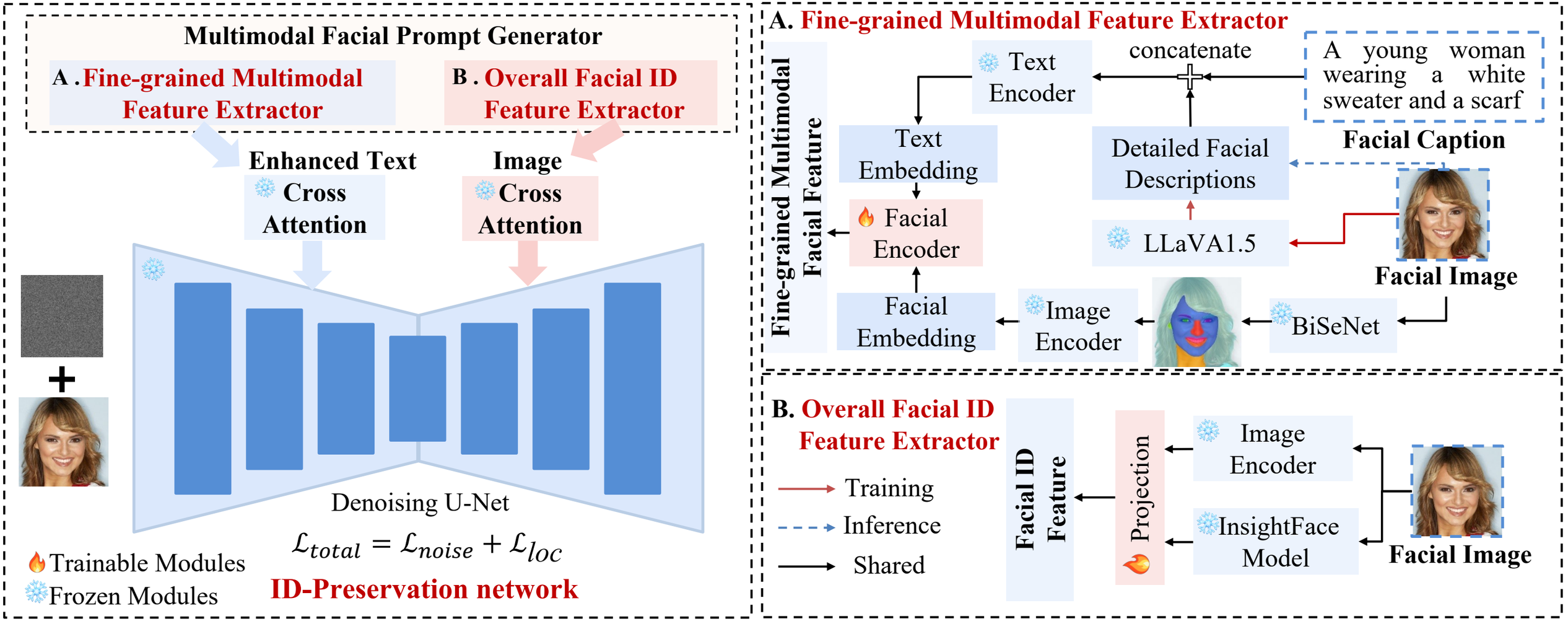
Multimodal Facial Prompt Generator
A . Fine-grained Multimodal Feature Extractor
B . Overall Facial ID Feature Extractor
Enhanced Text
Cross Attention
Image
Cross Attention
Denoising U-Net
$\mathcal{L}_{total} = \mathcal{L}_{noise} + \mathcal{L}_{loc}$
ID-Preservation network
Trainable Modules
Frozen Modules
A. Fine-grained Multimodal Feature Extractor
Fine-grained Multimodal Facial Feature
Text Encoder
concatenate
A young woman wearing a white sweater and a scarf
Facial Caption
Text Embedding
Detailed Facial Descriptions
Facial Encoder
LLaVA1.5
Facial Image
Facial Embedding
Image Encoder
BiSeNet
B. Overall Facial ID Feature Extractor
Training
Inference
Shared
Facial ID Feature
Projection
Image Encoder
InsightFace Model
Facial Image