

CLIP Images Projection Vision Encoder LLM Write a detailed description of the given image. VAE Encoder 7 CLIP Add noise & Text Encoder SD Unet Repeat until all prompt text is encoded $\varepsilon_{\theta}(\boldsymbol{Z}_{t},\boldsymbol{C})$ (b) Stage 2: Training Stable Diffusion adapted to LLM instructions Masks I_{mask} – Inpainting → l_{edit} SD $\mathcal{F}_{\text{edit}}$ Images -Generation I_{input} Multimoal LLM xxxxxxxxxxxxx Text T2I SD \mathcal{F}_{MLLM} $\rightarrow I_{gen}$ Instructions T_{input} (c) Inference process of LLMGA

 I_{input}

(a) Stage 1: Training multimoal LLM for image generation