```
#define PROMPT_TEXT_PREFIX "<|im_start|>system You are a helpful
assistant. <|im_end|> <|im_start|>user"
#define PROMPT_TEXT_POSTFIX "<|im_end|><|im_start|>assistant"

RKLLMInput rkllm_input;
string input_str = "<image>Please describe the image shortly.";
input_str = PROMPT_TEXT_PREFIX + input_str + PROMPT_TEXT_POSTFIX;
rkllm_input.multimodal_input.prompt = (char*)input_str.c_str();

rkllm_input.multimodal_input.n_image_tokens = 256;
int rkllm_input_len = multimodal_input.n_image_tokens * 3072;
rkllm_input.multimodal_input.image_embed = (float
*)malloc(rkllm_input_len * sizeof(float));
FILE *file;
file = fopen("models/image_embed.bin", "rb");
fread(rkllm_input.multimodal_input.image_embed, sizeof(float),
rkllm_input_len, file);
fclose(file);

rkllm_input.input_type = RKLLM_INPUT_MULTIMODAL;
RKLLMInferParam rkllm_infer_params;
memset(&rkllm_infer_params, 0, sizeof(RKLLMInferParam));
rkllm_infer_params.mode = RKLLM_INFER_GENERATE;
```

RKLLM supports different inference modes and defines the RKLLMInferParam structure. It currently supports joint inference with preloaded LoRA models during the inference process, or saving a Prompt Cache for subsequent inference acceleration. The specific definition is as follows:

Table 3-17 Explanation of RKLLMInferParam Structure

Definition	RKLLMInferParam
Introduction	Used to define different inference modes.
Struct Fields	***RKLLMInferMode mode:*** Inference mode, supporting RKLLM_INFER_GENERATE normal inference mode and RKLLM_INFER_GET_LOGITS additional logits retrieval inference mode. ***RKLLMLoraParam lora_params*:*** Parameter configuration for the LoRA used during inference, used to select which LoRA to infer when multiple LoRAs are loaded. Set to NULL if LoRA is not needed. ***RKLLMPromptCacheParam prompt_cache_params*:*** Parameter configuration for using the Prompt Cache during inference. Set to NULL if Prompt Cache generation is not needed.