(a) VLM-based Approach

Image Encoder

(Optional) Textual Inversion

Fusion Module

Text Encoder

Shows two women taking pictures from two different angles of a similarly sized and colored dog of a different breed.

Reference Image | Modification Instruction | Retrieved Image

(b) MLLM-based Approach

Multimodal Large Language Model

Shows two women taking pictures from two different angles of a similarly sized and colored dog of a different breed.

(c) Ours (CIR-CoT)

MLLM

Reasoning CoT

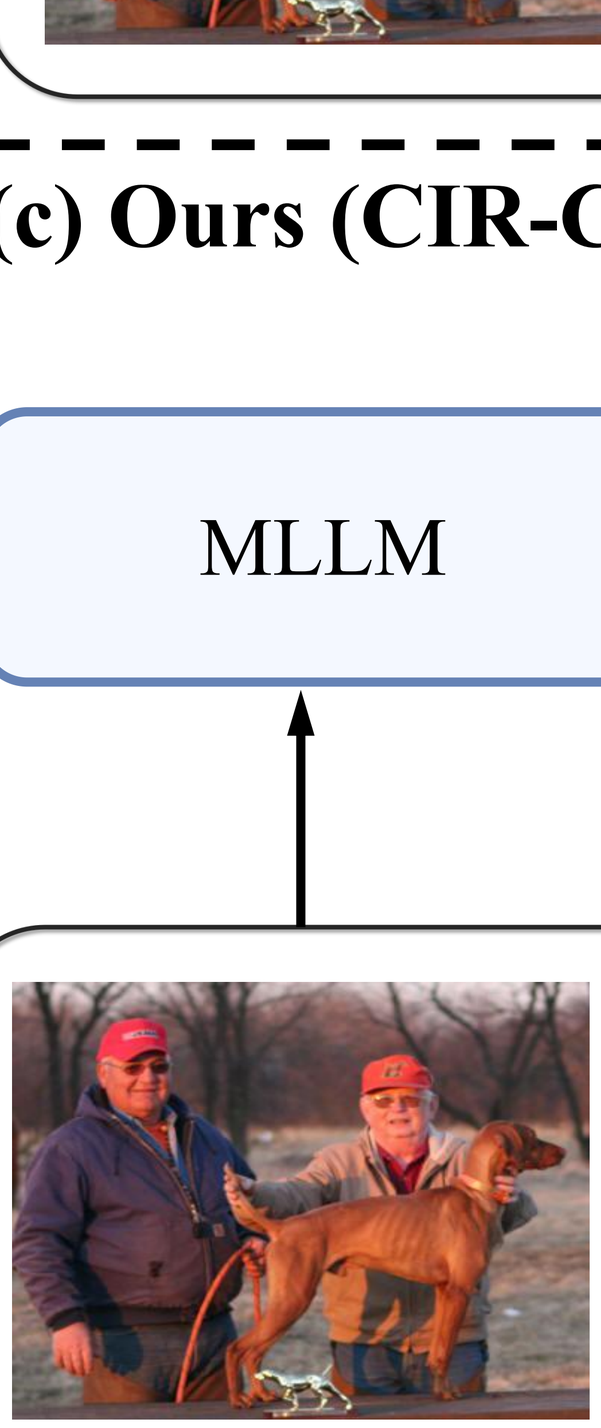
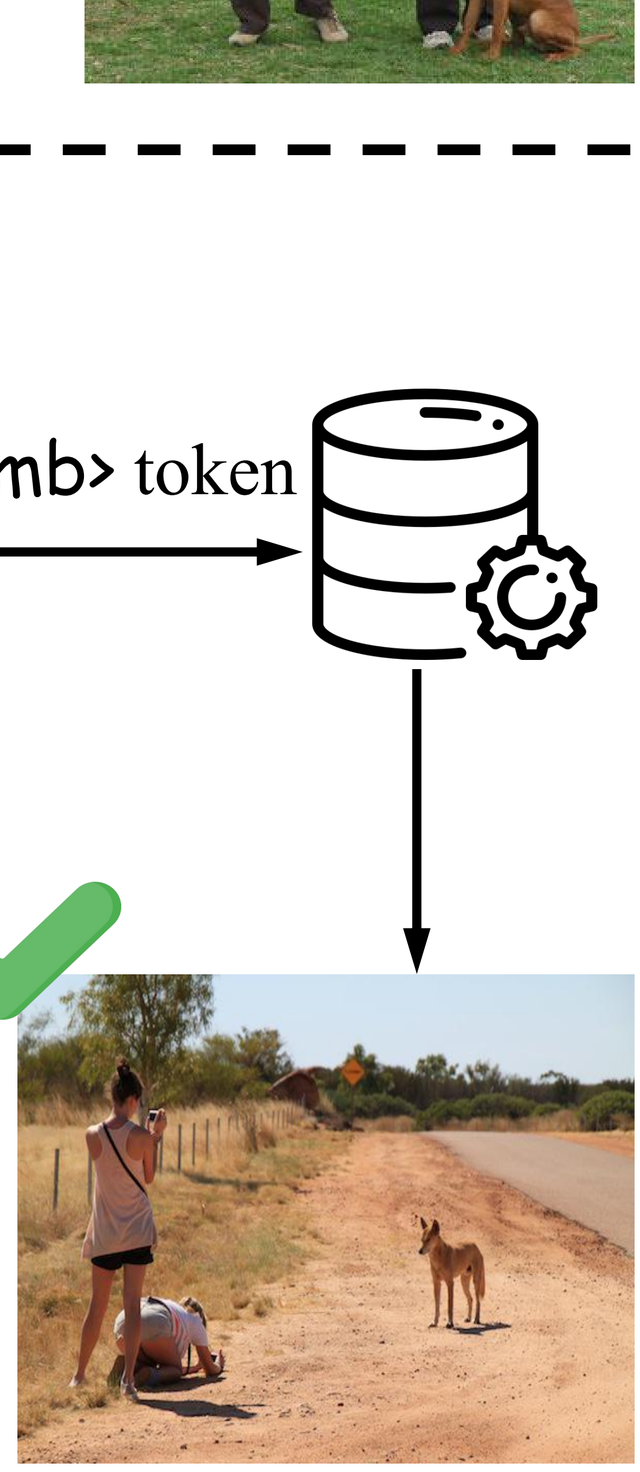
<Caption>.....</Caption>

<Reasoning> .....</Reasoning>

<Conclusion>.....</Conclusion>

The modified image can be summarized as <emb>.

<emb> token



Shows two women taking pictures from two different angles of a similarly sized and colored dog of a different breed.